b)

(19)日本国特許庁（ＪＰ）　(12)**公開特許公報**（Ａ）　(11)特許出願公開番号
特開2000－101442
(P2000－101442A)
(43)公開日　平成12年４月７日(2000.4.7)

(51)Int.Cl.⁷	識別記号	ＦＩ		テーマコード（参考）
Ｈ０３Ｍ　7/40		Ｈ０３Ｍ　7/40		
Ｇ０６Ｆ　5/00		Ｃ０６Ｆ　5/00	Ｈ	

審査請求　未請求　請求項の数26　ＯＬ　（全 31 頁）

(21)出願番号　特願平10－272724

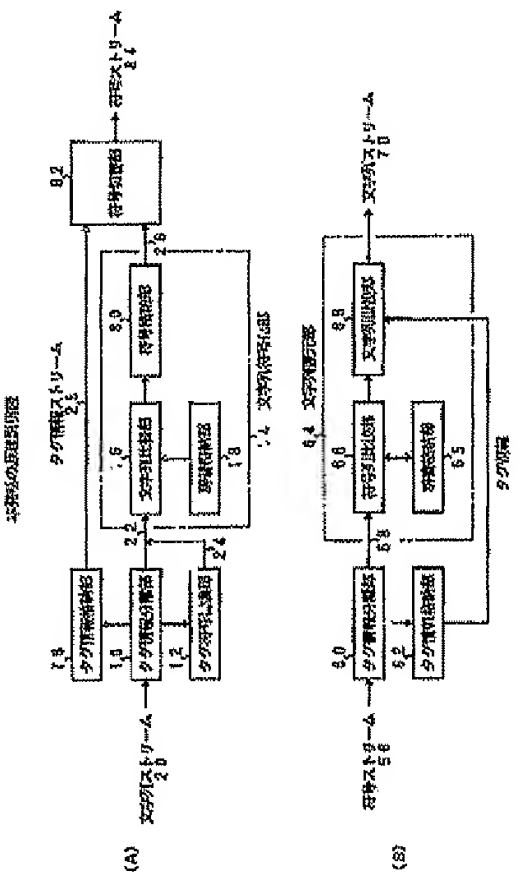
(22)出願日　平成10年９月28日(1998.9.28)

(71)出願人　000005223
富士通株式会社
神奈川県川崎市中原区上小田中４丁目１番１号
(72)発明者　森原　隆
神奈川県川崎市中原区上小田中４丁目１番１号　富士通株式会社内
(72)発明者　矢作　裕紀
神奈川県川崎市中原区上小田中４丁目１番１号　富士通株式会社内
(74)代理人　100079359
弁理士　竹内　進　（外１名）

最終頁に続く

(54)【発明の名称】　データ圧縮装置及び復元装置並びにその方法

(57)【要約】
【課題】タグを含む構造化文書を、文書の検索あるいは読込み時間の短縮とメモリやディスク容量の増加を最小限とするように圧縮し復元する。
【解決手段】タグ情報分離部１０は、文字列ストリームから識別したタグを分離してタグ情報として出力する。タグ置換部１２は、タグが分離された文字列ストリームの位置に識別のためにタグ符号を配置する。文字列符号化部１４は、タグ符号置換部１２から出力されるタグ符号を含む文字列ストリームを圧縮符号化する。データ復元装置は、圧縮と逆の操作で文字列ストリームを復元する。



【特許請求の範囲】
【請求項１】タグを含む文書で構成された文字列ストリームから符号データを生成するデータ圧縮装置に於いて、
前記文字列ストリームから識別したタグを分離してタグ情報として出力するタグ情報分離部と、
前記タグ情報分離部でタグが分離された文字列ストリームの位置に識別のためにタグ符号を配置するタグ符号置換部と、
前記タグ符号置換部から出力されたタグ符号を含む文字列ストリームを符号化して符号ストリームを出力する文字列符号化部と、を有することを特徴とするデータ圧縮装置。
【請求項２】請求項１記載のデータ圧縮装置に於いて、前記タグ符号置換部は、タグが分離された文字列ストリームの位置に、所定の固定符号を前記タグ符号として配置することを特徴とするデータ圧縮装置。
【請求項３】請求項１記載のデータ圧縮装置に於いて、前記タグ符号置換部は、タグが分離された文字列ストリームの位置に、前記タグ情報分離部で分離されたタグの出現順序を示すタグ符号を配置することを特徴とするデータ圧縮装置。
【請求項４】請求項１記載のデータ圧縮装置に於いて、更に、
前記タグ情報分離部で分離されたタグ情報を格納するタグ情報格納部と、
前記文字列符号化部で生成された符号データを格納する符号格納部と、
前記タグ情報格納部に格納されたタグ情報と符号格納部に格納された符号データを選択して出力する符号切替部と、を設けたことを特徴とするデータ圧縮装置。
【請求項５】請求項１記載のデータ圧縮装置に於いて、前記文字列符号化部は、
圧縮する際の処理単位となる文字列を登録した辞書を格納する辞書格納部と、
前記タグ符号置換部からの文字列ストリームの中の部分文字列と前記辞書格納部の登録文字列との比較により、前記登録文字列に一致する部分文字列を検出し、検出した部分文字列ごとに予め定められた符号を割り当てて出力する文字列比較部と、を備えたことを特徴とするデータ圧縮装置。
【請求項６】請求項１記載のデータ圧縮装置に於いて、更に、前記タグ情報分離部で分離したタグ情報を圧縮するタグ情報圧縮部を設けたことを特徴とするデータ圧縮装置。
【請求項７】請求項１記載のデータ圧縮装置に於いて、更に、
圧縮する際の処理単位となるタグ情報中のタグ文字列を登録した辞書を格納するタグ辞書格納部と、
前記タグ情報分離部で分離したタグ情報に含まれる文字列ストリームの部分文字列と前記タグ辞書格納部の登録文字との比較により、前記登録文字列に一致する部分文字列を検出し、検出した部分文字列ごとに予め定められた符号を割り当てて出力するタグ文字列比較部と、を備えたことを特徴とするデータ圧縮装置。
【請求項８】請求項４記載のデータ圧縮装置に於いて、更に、前記文字列符号化部で生成した符号データの中のタグ位置を検出するタグ位置検出部を設け、前記タグ情報格納部に前記タグ情報分離部で分離したタグ情報と共に前記タグ位置検出部で検出したタグ位置の指定情報を格納したことを特徴とするデータ圧縮装置。
【請求項９】請求項８記載のデータ圧縮装置に於いて、前記タグ位置検出部は、文書先頭又は特定のタグからの符号量を検出して前記タグ情報格納部にタグ情報と共に格納したことを特徴とするデータ圧縮装置。
【請求項１０】タグを含む文書の文字列ストリームから分離したタグ情報と、分離したタグの位置にタグ符号を配置した文字列ストリームを符号化した符号データとを含む符号ストリームから文字列データを復元するデータ復元装置に於いて、
前記符号ストリームからタグ情報と符号データとを分離するタグ情報分離部と、　前記タグ情報分離部で分離したタグ情報を格納するタグ情報格納部と、
前記符号データから文字列及びタグ符号を含む文字列データを復元した後に、前記タグ符号をダク情報格納部のタグ情報に置き換える文字列復元部と、を備えたことを特徴とするデータ復元装置。
【請求項１１】請求項１０記載のデータ復元装置に於いて、前記文字列復元部は、
復元する際の処理単位となる文字列の符号に対応した復元文字列を登録した辞書を格納する辞書格納部と、
前記符号ストリームから復元単位となる文字列の符号を分離して前記辞書格納部の参照で元の文字列を復元する文字列比較部と、
前記文字列比較部により復元したタグ符号を、前記ダク情報格納部のタグ情報に置き換える文字列置換部と、を備えたことを特徴とするデータ復元装置。
【請求項１２】請求項１０記載のデータ復元装置に於いて、更に、前記タグ情報格納部に格納されたタグ情報の圧縮データを復元するタグ情報復元部を設けたことを特徴とするデータ復元装置。
【請求項１３】請求項１０記載のデータ復元装置に於いて、更に、
復元する際の処理単位となるタグ文字列の符号に対応した復元文字列を登録した辞書を格納するタグ辞書格納部と、
前記タグ情報分離部により分離したタグ情報から復元単位となるタグ文字列の符号を分離し、前記辞書格納部の参照で元のタグ文字列を復元するタグ文字列比較部と、を備えたことを特徴とするデータ復元装置。

【請求項１４】タグを含む文書で構成された文字列ストリームから符号データを生成するデータ圧縮方法に於いて、
前記文字列ストリームから識別したタグを分離してタグ情報として出力するタグ情報分離過程と、
前記タグ情報分離過程でタグが分離された文字列ストリームの位置に識別のためにタグ符号を配置するタグ符号置換過程と、
前記タグ符号置換過程から出力されたタグ符号を含む文字列ストリームを符号化して符号ストリームを出力する文字列符号化過程と、を有することを特徴とするデータ圧縮方法。
【請求項１５】請求項１４記載のデータ圧縮方法に於いて、前記タグ符号置換過程は、タグが分離された文字列ストリームの位置に、所定の固定符号を前記タグ符号として配置することを特徴とするデータ圧縮方法。
【請求項１６】請求項１４記載のデータ圧縮方法に於いて、前記タグ符号置換過程は、タグが分離された文字列ストリームの位置に、前記タグ情報分離過程で分離されたタグの出現順序を示すタグ符号を配置することを特徴とするデータ圧縮方法。
【請求項１７】請求項１４記載のデータ圧縮方法に於いて、更に、
前記タグ情報分離過程で分離されたタグ情報を格納するタグ情報格納過程と、
前記文字列符号化過程で生成された符号データを格納する符号格納過程と、
前記タグ情報格納過程に格納されたタグ情報と符号格納過程に格納された符号データを選択して出力する符号切替過程と、を設けたことを特徴とするデータ圧縮方法。
【請求項１８】請求項１４記載のデータ圧縮方法に於いて、前記文字列符号化過程は、
圧縮する際の処理単位となる文字列を登録した辞書を生成する辞書生成過程と、
前記タグ符号置換過程で得られた文字列ストリームの中の部分文字列と前記辞書の登録文字列との比較により、前記登録文字列に一致する部分文字列を検出し、検出した部分文字列ごとに予め定められた符号を割り当てて出力する文字列比較過程と、を備えたことを特徴とするデータ圧縮方法。
【請求項１９】請求項１４記載のデータ圧縮方法に於いて、更に、前記タグ情報分離過程で分離したタグ情報を圧縮するタグ情報圧縮過程を設けたことを特徴とするデータ圧縮方法。
【請求項２０】請求項１４記載のデータ圧縮方法に於いて、更に、
圧縮する際の処理単位となるタグ情報中のタグ文字列を登録した辞書を生成するタグ辞書生成過程と、
前記タグ情報分離過程で分離したタグ情報に含まれる文字列ストリームの部分文字列と前記タグ辞書の登録文字との比較により、前記登録文字列に一致する部分文字列を検出し、検出した部分文字列ごとに予め定められた符号を割り当てて出力するタグ文字列比較過程と、を備えたことを特徴とするデータ圧縮方法。
【請求項２１】請求項１７記載のデータ圧縮方法に於いて、更に、前記文字列符号化過程で生成した符号データのタグ位置を検出するタグ位置検出過程を設け、前記タグ情報分離過程で分離したタグ情報と共に前記タグ位置検出過程で検出したタグ位置の指定情報を格納したことを特徴とするデータ圧縮方法。
【請求項２２】請求項２１記載のデータ圧縮方法に於いて、前記タグ位置検出過程は、文書先頭又は特定のタグからの符号量を検出して前記タグ情報格納過程で分離したタグ情報と共に格納することを特徴とするデータ圧縮方法。
【請求項２３】タグを含む文書の文字列ストリームから分離したタグ情報と、分離したタグの位置にタグ符号を配置した文字列ストリームを符号化した符号データとを含む符号ストリームから文字列データを復元するデータ復元方法に於いて、
前記タグ情報と符号データとを分離するタグ情報分離過程と、
前記タグ情報分離過程で分離したタグ情報を格納するタグ情報格納過程と、
前記符号データから文字列及びタグ符号を含む文字列ストリームを復元した後に、前記タグ符号を前記ダク情報格納過程で分離したタグ情報に置き換える文字列復元過程と、を備えたことを特徴とするデータ復元方法。
【請求項２４】請求項２３記載のデータ復元方法に於いて、前記文字列復元過程は、
復元する際の処理単位となる文字列の符号に対応した復元文字列を登録した辞書を生成する辞書生成過程と、
前記符号ストリームから復元単位となる文字列の符号を分離して前記辞書の参照で元の文字列を復元する文字列比較過程と、
前記文字列比較過程により復元したタグ符号を、前記ダク情報格納過程で分離したタグ情報に置き換える文字列置換過程と、を備えたことを特徴とするデータ復元方法。
【請求項２５】請求項２３記載のデータ復元方法に於いて、更に、前記タグ情報格納過程で格納されたタグ情報の圧縮データを復元するタグ情報復元過程を設けたことを特徴とするデータ復元方法。
【請求項２６】請求項２３記載のデータ復元方法に於いて、更に、
復元する際の処理単位となるタグ文字列の符号に対応した復元文字列を登録した辞書を生成するタグ辞書生成過程と、
前記タグ情報分離過程により分離したタグ情報から復元単位となるタグ文字列の符号を分離し、前記辞書の参照

で元のタグ文字列を復元するタグ文字列比較過程と、を備えたことを特徴とするデータ復元方法。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、タグを含む構造化文書で構成された文字列ストリームから符号データを生成するデータ圧縮装置及び復元装置並びにその方法に関し、特に構造化文書の文字列ストリームからタグ情報を分離して符号化と復元を行うためのデータ圧縮装置及び復元装置並びにその方法に関する。

【０００２】

【従来の技術】近年、文字コード、画像データ等の様々な種類のデータがコンピュータで扱われている。さらに、インターネット・イントラネットの普及に伴い、電子メールや電子化文書が増加している。このような大量のデータは、データ中の冗長な部分を省いて圧縮することにより、記憶容量を減らしたり、短時間で遠隔地に送ることを可能にしている。

【０００３】本発明の分野は、文字コードの圧縮に限らず、様々なデータに適用できるが、以下では、情報理論で用いられる呼称を踏襲し、データの１ワード単位を文字と呼び、データが任意のワードつながったものを文字列と呼ぶことにする。

【０００４】最近では、コンピュータ上で扱う文書の形式を統一する動きがある。その中で、文書の作成を効率良く行うため、タグを用いて文書内容を部分的に区別して、予め見出しや段落などの複数の文書部品を作成し、各々の文書部品間の関係を定めて文書を構造化して編集することが試みられている。

【０００５】このような文書に構造の概念を取り入れた構造化文書の例としては、国際規格のＯＤＡ（ISO ８６１３：Open Document Architecture）や、ＳＧＭＬ（ISO8879：Standard Markup Language）の規格による構造化文書がある。またこのような構造文書を用いた文書処理方法は、例えば特開平５－１３５０５４号のものがある。

【０００６】ＳＧＭＬによる構造化文書は、従来のテキスト処理システムとの親和性が高く、米国を中心に普及し、実用化されてきている。ＳＧＭＬによる構造化文書は、予め文書構造の雛型が与えられ、文書構造は雛型の範囲に制限される。

【０００７】図２５はＳＧＭＬの構造化文書であり、ＳＧＭＬ宣言２００、文書型定義（ＤＴＤ：Document Type Definition）２０２、及び文書実現値２０４の３つの部分からなる。このうち文書の構造を定義する雛型が文書型定義２０２であり、図２６のように、章、節、タイトルなどの文書構造を定義している。

【０００８】ＳＧＭＬの構造化文書では、文書構造を表現するために、文書テキスト内にタグと呼ばれる識別子を用いて、文書テキストを区分する。図２７はＳＧＭＬの構造化文書の具体例であり、例えば文書のタイトルの場合、「<TITLE> 発明（考案）明細書</TITLE>」で表現される。即ち、開始タグである「<TITLE> 」と終了タグである「</TITLE>」で囲まれた文字が要素であり、この場合はタイトル内容「発明（考案）明細書」を表わす。

【０００９】現在、公的機関を中心にＳＧＭＬを採用する例が増えてきている。特に米国では、国防総省が文書をＳＧＭＬで記述して納入することを義務づけている。日本においても特許庁のＣＤ－ＲＯＭ公報として、この構造化文書を採用している。また、インターネットで使われているＷＷＷ（World Wide Web）の記述形式として普及しているＨＴＭＬ（Hyper Text Markup Language）は、ＳＧＭＬの一形態である。

【００１０】このようなＳＧＭＬ等の構造化文書を圧縮する方法として、本願出願人は、特開平９－２６１０７２号の方法を提案している。

【００１１】この方法では、タグ情報を有する構造化文書の文書データが入力された場合、文書型定義ＤＴＤなどで定義されているタグ情報を検出する。タグ情報が検出された場合、タグ情報は何ら変換せずに、そのまま出力する。さらにタグ情報を検出したことにより、タグ情報以外の入力文字列を符号化するモードに移行する。

【００１２】この符号化の基本アルゴリズムは図２８のようになる。まずステップＳ１で入力された文字又は文字列が予め登録した辞書の文字又は文字列と同一か否か検索して比較し、同一であればステップＳ２で入力データを辞書の登録番号で符号化し、ステップＳ３で符号を出力する。

【００１３】ステップＳ１で同一の登録文字又は文字列が検索できなかった場合は、ステップＳ５で元の入力文字又は文字列をそのまま出力する。このような処理をステップＳ５で入力文字列がなくなるまで繰り返す。

【００１４】図２７のＳＧＭＬ文書ファイルについて図２８の符号化を行うと、図２９の圧縮データファイルが得られる。この圧縮データファイルは、１つのファイル中に圧縮されていないタグ情報の部分と圧縮されたテキスト文書の部分とが混在する形式となる。

【００１５】

【発明が解決しようとする課題】文書テキストを圧縮する方法は、膨大なデータ量である文書テキストを実用に耐え得るデータ量に圧縮することができ、電子化された文書テキストを実現する上で非常に有用な技術である。

【００１６】しかしながら、図２９のような構造化文書の圧縮データファイルにあっては、ファイル中のタグ情報を検索する場合、タグ情報は圧縮されない部分として圧縮された文書データの中に混在しており、ファイル全体をメモリ上に展開して必要とするタグ情報を検索しなければならない。また圧縮部分となる本文中のキーワードを検索したい場合にも、同様にファイル全体をメモリ上に展開して処理する必要がある。

【００１７】このため、構造化文書の圧縮データファイルから必要とする文書を検索あるいは入手するために、文書としては不要な部分の読み込みが必要となり、データ伝送量が増加して読込みに時間がかかり、また大きなメモリ領域とディスク容量の確保が必要となる問題がある。

【００１８】本発明の目的は、タグ情報を含む構造化文書の圧縮データにつき、文書の検索あるいは読込み時間の短縮とメモリやディスク容量の増加を最小限とするためのデータ圧縮装置及び復元装置並びにその方法を提供することにある。

【００１９】

【課題を解決するための手段】図１は本発明の原理説明図である。

（圧縮）まず本発明は、図１（Ａ）のように、タグを含む文書で構成された文字列ストリーム２０から符号データを生成するデータ圧縮装置を対象とする。このデータ圧縮装置として本発明は、文字列ストリームから識別したタグを分離してタグ情報として出力するタグ情報分離部１０と、タグ情報分離部１０でタグが分離された文字列ストリームの位置に識別のためにタグ符号２４を配置するタグ符号置換部１２と、タグ符号置換部１２から出力されたタグ符号を含む文字列ストリーム２２を符号化して符号ストリーム８４を出力する文字列符号化部１４とを設けたことを特徴とする。

【００２０】このような本発明のデータ圧縮装置によれば、タグを含む構造化文書の文字列ストリームにつき、タグ情報と本文（文字列）とを分離し、少なくとも本文を符号化することで高い圧縮率を実現し、分離したタグ情報を検索することで、検索が高速化できる。

【００２１】例えば、圧縮データファイルの中の本文から分離されたタグ情報を検索し、一致するタグ情報が検索できたら、復元した本文中のタグ符号を検索したタグ情報までの数だけ読み飛ばすことで、目標とする文書の先頭に容易に到達することができる。

【００２２】タグ符号置換部１２は、タグが分離された文字列ストリームの位置に、所定の固定符号をタグ符号として配置する。タグ符号として固定符号を使用することで、本文中のタグ位置の検索が簡単にできる。

【００２３】またタグ符号置換部１２は、タグが分離された文字列ストリームの位置に、タグ情報分離部１０で分離されたタグの出現順序を示すタグ符号を配置する。このようにタグ符号に出現順序の情報を持たせることで、タグ情報に基づく本文検索の高速化と信頼性が高められる。

【００２４】データ圧縮装置は、更に、タグ情報分離部１０で分離されたタグ情報を格納するタグ情報格納部７８と、文字列符号化部１４で生成された符号データを格納する符号格納部８０と、タグ情報格納部７８に格納されたタグ情報と符号格納部８０に格納された符号データを選択して出力する符号切替部８２と設ける。このように分離したタグ情報と本文の符号データを個別に格納することで、圧縮データの検索や転送要求に対する管理を容易にする。

【００２５】文字列符号化部１４は、圧縮する際の処理単位となる文字列を登録した辞書を格納する辞書格納部１８と、タグ符号置換部１２からの文字列ストリーム２２の中の部分文字列と辞書格納部１８の登録文字列との比較により、登録文字列に一致する部分文字列を検出し、検出した部分文字列ごとに予め定められた符号を割り当てて出力する文字列比較部１６とを備える。

【００２６】この文字列符号化部１４による符号化処理は、スペースで区切られない単語構造をもつ言語の文字コードで作成された文書データの圧縮に効果があり、スペースで区切られない単語構造をもつ言語としては、例えば日本語、中国語、ハングル語等がある。

【００２７】日本語を例にとると、日本語の単語の関する（株）日本電子化辞書研究所（ＥＤＲ）の研究成果がある（横井，木村，小泉，三吉、「表層レベルにおける電子化辞書の情報構造」、情報処理学会論文誌，Vol.37，No.3，p.333-344，1996 ）。

【００２８】この研究結果では、日本語を構成する形態素、即ち単語の品詞を集計している。単純に単語を品詞類に分けて登録すると１３６，４８６個となり、１７ビット（最大２６２，１４３個）の符号で表わすことができる。

【００２９】また新世代コンピュータ技術開発機構（ＩＣＯＴ）で作成した日本語単語辞書を構成する約１３万語の単語ごとに構成する文字数を検出し、その分布を求めた結果、全登録単語の１／２以上の７万語が２文字で構成され、平均文字数は２．８文字（４４．８ビット）であることも判明している。

【００３０】図１（Ａ）の辞書格納部１８は、日本語の辞書として実用的な例えば約１３万語の単語を、各単語に例えば１７ビットの固定長の文字列符号を割り当てた辞書を作成して格納し、非圧縮データの部分文字列に一致する辞書の登録文字列を検索して１７ビットの固定長符号を文字列符号として割り当てて出力することで、文書データの大小に関わらず、実質的にデータ量を半分以下に圧縮することができる。

【００３１】本発明のデータ圧縮装置は、タグ情報分離部１０で分離したタグ情報を圧縮するタグ情報圧縮部を設ける。タグ情報は、単独のタグと、タグと文字列の組合せを含むが、このタグ情報圧縮部は、タグと文字列を区別することなく一括してタグ情報を圧縮する。この圧縮は例えばＬＺ７７、ＬＺ７８、算術符号化等のアルゴリズムを使う。

【００３２】本発明のデータ圧縮装置は、タグ情報中の日本語等のスペースで区切られない言語の文字列を対象に、本文の文字列符号化部と同じ符号化を行ってタグ情

報を圧縮する。即ち、本発明のデータ圧縮装置は、圧縮する際の処理単位となるタグ情報中のタグ文字列を登録した辞書を格納するタグ辞書格納部と、タグ情報分離部１０で分離したタグ情報に含まれる文字列ストリームの部分文字列とタグ辞書格納部の登録文字との比較により、登録文字列に一致する部分文字列を検出し、検出した部分文字列ごとに予め定められた符号を割り当てて出力するタグ文字列比較部とを備えたことを特徴とする。

【００３３】このように分離したタグ情報についても圧縮することで、文字列符号化部１４による本文の圧縮と併せて文書ファイル全体を高圧縮できる。

【００３４】本発明のデータ圧縮装置に於いて、更に、文字列符号化部１４で生成した符号データのタグ位置を検出するタグ位置検出部を設け、タグ情報格納部７８にタグ情報分離部１０で分離したタグ情報と共にタグ位置検出部で検出したタグ位置の指定情報を格納する。この場合、タグ位置検出部は、文書先頭又は特定のタグからの符号量を検出してタグ情報格納部にタグ情報と共に格納する。

【００３５】このように分離したタグ情報に、圧縮した本文の対応するタグ符号の位置を示す文書先頭又は特定タグからのデータ量（バイト数）が位置指定情報として格納されているため、タグ情報から必要とするタグを検索したい場合、直ちに本文に圧縮データ中の対応するタグ符号の位置が特定でき、必要とする本文のランダムアクセスが効率良くできる。

【００３６】（復元）本発明は、タグを含む文書の文字列ストリームから分離したタグ情報と、分離したタグの位置にタグ符号を配置した文字列ストリームを符号化した符号データとを含む符号ストリームから文字列データを復元するデータ復元装置を対象とする。

【００３７】このデータ復元装置として本発明は、図１（Ｂ）のように、符号ストリーム５６からタグ情報と符号データとを分離するタグ情報分離部６０と、タグ情報分離部６０で分離したタグ情報を格納するタグ情報格納部６２と、符号データから文字列及びタグ符号を復元した後に、タグ符号をダク情報格納部６２のタグ情報に置き換える文字列復元部６４とを備えたことを特徴とする。

【００３８】文字列復元部６４は、図１（Ａ）の文字列符号化部１４の逆の操作を行うもので、復元する際の処理単位となる文字列の符号に対応した復元文字列を登録した辞書を格納する辞書格納部６５と、符号ストリームから復元単位となる文字列の符号を分離して辞書格納部６５の参照で元の文字列を復元する文字列比較部６６と、文字列比較部６６により復元したタグ符号をダク情報格納部６８のタグ情報に置き換える文字列置換部６８とを備える。

【００３９】本発明のデータ復元装置は、データ圧縮装置側でタグ情報をＬＺ７７，ＬＬＺ７８等で圧縮している場合、タグ情報格納部６２に格納されたタグ情報の圧縮データを復元するタグ情報復元部を設ける。

【００４０】また本発明のデータ復元装置は、データ圧縮装置側でタグ情報の文字列を符号化している場合、復元する際の処理単位となるタグ文字列の符号に対応した復元文字列を登録した辞書を格納するタグ辞書格納部と、タグ情報分離部６０により分離したタグ情報から復元単位となるタグ文字列の符号を分離し、タグ辞書格納部の参照で元のタグ文字列を復元するタグ文字列比較部とを備える。

【００４１】本発明は、更に、タグ情報を含む構造化文書の圧縮方法及び復元方法を提供する。本発明によるタグを含む文書で構成された文字列ストリームから符号データを生成するデータ圧縮方法は、文字列ストリームから識別したタグを分離してタグ情報として出力するタグ情報分離過程と、タグ情報分離過程でタグが分離された文字列ストリームの位置に識別のためにタグ符号を配置するタグ符号置換過程と、タグ符号置換過程から出力されたタグ符号を含む文字列ストリームを符号化して符号ストリームを出力する文字列符号化過程と、を備える。

【００４２】また本発明によるタグを含む文書の文字列ストリームから分離したタグ情報と、分離したタグの位置にタグ符号を配置した文字列ストリームを符号化した符号データとを含む符号ストリームから文字列データを復元するデータ復元方法を提供する。この復元方法は、タグ情報と符号データとを分離するタグ情報分離過程と、タグ情報分離過程で分離したタグ情報を格納するタグ情報格納過程と、符号データから文字列及びタグ符号を復元した後に、タグ符号をダク情報格納過程で分離したタグ情報に置き換える文字列復元過程と、を備える。データ圧縮方法及び復元方法の詳細は、装置の場合と同じになる。

【００４３】

【発明の実施の形態】図２は本発明のデータ圧縮装置の第１実施形態のブロック図である。図２において、第１実施形態のデータ圧縮装置は、タグ情報分離部１０、タグ符号置換部１２及び文字列符号化部１４で構成される。文字列符号化部１４には文字列比較部１６と辞書格納部１８が設けられる。

【００４４】タグ情報分離部１０は、例えば図２７に示したＳＧＭＬ日本語文書ファイルから読み出した文字列ストリーム２０を入力し、入力した文字列ストリーム２０に含まれるタグを識別し、識別したタグを分離してタグ情報ストリーム２８として出力する。

【００４５】タグ符号置換部１２は、タグ情報分離部１０でタグ情報が分離された文字列ストリームのタグ位置に予め定めたタグ符号を配置し、タグ符号配置済みの文字列ストリーム２２を文字列符号化部１４に供給する。文字列符号化部１４はタグ符号置換部１２により配置されたタグ符号を含む文字列ストリーム２２を符号化し、

符号ストリーム２６を出力する。

【００４６】図３は、図２のタグ情報分離部１０の詳細であり、タグ符号置換部１２と共に示している。タグ情報分離部１０は、タグ比較部３０、タグ識別規則格納部３２及び出力切替部３４で構成される。タグ識別規則格納部３２にはＳＧＭＬ文書における文書型定義ＤＴＤから得られたタグ情報の識別規則が格納されている。

【００４７】タグ比較部３０は文字列ストリーム２０を入力し、タグ比較部３０でタグ識別規則格納部３２の識別規則と比較し、タグ情報識別により比較出力が得られると、出力切替部３４を文字列ストリーム２２の出力からタグ情報ストリーム２８の出力に切り替え、識別したタグ情報をタグ情報ストリーム２８として出力する。

【００４８】同時にタグ符号置換部１２にタグ情報識別に基づく比較出力を行い、タグ符号置換部１２に予め設定されているタグ符号２４を、出力切替部３４から出力が断たれたタグ情報の位置に挿入配置する。タグ符号置換部１２により文字列ストリーム２２のタグ情報の位置に配置されるタグ情報２４としては、例えば１６進の固定符号「０ｘ００００」を使用する。

【００４９】図４は、ＳＧＭＬ日本語文書ファイルから読み出した文字列ストリーム２０の図２のデータ圧縮装置による圧縮処理の説明図である。

【００５０】図２のタグ情報分離部１０に対する文字列ストリーム２０として入力されるＳＧＭＬ日本語文書ファイル３５は、図３のタグ情報分離部１０に設けているタグ比較部３０でタグ識別規則格納部３２に格納しているタグ識別規則と比較され、例えば先頭の「＜ＴＩＴＬＥ＞発明（考案）の明細書＜／ＴＩＴＬＥ＞」がタグ情報として識別され、このタグ情報はタグ情報ファイル３６の先頭位置のように分離される。

【００５１】またタグ情報の分離と並行して、ＳＧＭＬ日本語文書ファイル３５のタグ情報を分離した位置に、１６進の固定符号「０ｘ００００」を用いたタグ符号が挿入配置され、このタグ情報のタグ符号への置換によりタグ置換済み日本語文書ファイル３８の文字列ストリームが生成される。

【００５２】分離されたタグ情報ファイル３６の内容となるタグ情報ストリームはそのまま出力される。またタグ置換済み日本語文書ファイル３８の内容となる文字列ストリームは、文字列符号化部１４により符号化されて符号ストリーム２６として出力される。

【００５３】図５は、図２７のＳＧＭＬ日本語文書ファイルの文字列ストリーム２０を図２のデータ圧縮装置に入力して、タグ符号置換部１２によりタグ情報を固定タグ符号に置換して得たタグ置換済み日本語文書ファイル３８である。このタグ置換済み日本語文書ファイルにあっては、図２７のＳＧＭＬ日本語文書ファイルにおけるタグ情報がそれぞれ「（タグ符号）」に置き換えられている。

【００５４】図６は図２７に示すＳＧＭＬ日本語文書ファイルの文字列ストリームから分離したタグ情報のタグ情報ファイル３６である。このタグ情報ファイル３６には、入力した文字列ストリームに含まれているタグ情報が順番に分離されて格納されている。

【００５５】図５のタグ置換済み日本語文書ファイル３８の内容となるタグ置換済みの文字列ストリーム２２は、図２の文字列符号化部１４で符号化され、圧縮された符号ストリーム２６として出力される。

【００５６】図７は、タグ符号としてタグ情報の出現順序を示す順序タグ符号を使用した場合のタグ置換済み日本語文書ファイル３８である。このタグ情報の出現頻度を表わす順序タグ符号としては、例えばタグの出現順序に応じて１６進で「０ｘ００１，０ｘ００２，０ｘ００３，・・・」等のように、一義に対応する順序タグ符号を使用すればよい。

【００５７】この出現順番を示す順序タグ符号を使用した場合には、図７のように日本語文字列データの中に置換されたタグ符号自体が「（タグ符号１），（タグ符号２），（タグ符号３），・・・」のように、文書先頭からの出現順序を表わしている。このため図６のように分離されたタグ情報の検索で図７の文書ファイル中の対応するタグ符号の位置を特定する際に、本文中の検索位置を簡単且つ確実に特定できる。

【００５８】例えば図６で５行目のタグ情報「＜ＳＥＣＴＩＯＮ＞請求項目の範囲＜／ＳＥＣＴＩＯＮ＞」の文書ファイル中の位置を知りたい場合には、このタグ識別情報は先頭から５番目に出現していることから、出現順序が５番目となる「（タグ符号５）」の位置を検索することで、簡単に特定できる。

【００５９】図８は、図２のデータ圧縮装置による圧縮処理のフローチャートである。まずステップＳ１で、入力文書の文字列ストリーム２０からタグ情報分離部１０によってタグ情報を分離して出力する。続いてステップＳ２で、入力文書の文字列ストリーム２０中のタグのあった位置にタグ符号置換部１２によって識別のためのタグ符号を挿入する。

【００６０】続いてステップＳ３で、タグ配置済みの文字列ストリーム中の文字列を文字列符号化部１４に設けている文字列比較部１６で辞書格納部１８内の対応する登録番号を符号として割り当て、符号ストリーム２６を出力する。このステップＳ１～Ｓ３の処理を、ステップＳ４で文字列ストリームの入力が終了するまで繰り返す。

【００６１】次に図２の文字列符号化部１４に設けた文字列比較部１６と、辞書格納部１８によるタグ置換済み文字列ストリーム２２の符号化処理を説明する。

【００６２】図２の文字列符号化部１４に設けた文字列比較部１６は、辞書格納部１５の参照により、単語を構成する文字列ごとに予め定めた所定の文字列符号を割り

当てる符号化を行う。

【００６３】まず文字列比較部１６で圧縮対象とする文書データとして、例えば日本語文書データを例にとると、日本語文書データの場合、１文字は２バイトのワードデータで構成されており、文書中の単語はスペースで区切られない構造を持っている。また日本語文書データは、１回の圧縮に使用する文書単位に入力しており、これはキロバイトオーダからメガバイトオーダの適宜のサイズの文書が入力される。

【００６４】文字列比較部１６は日本語文書データの文字列を先頭から順番に入力し、辞書格納部１５に予め登録されている単語単位の登録文字列と一致するか否かを検出する。文字列比較部１６で入力文字列に一致する登録文字列が検出されると、辞書格納部１５の一致検出された登録文字列に対応して予め登録されている文字列符号を読み出して割り当て、この文字列符号を文字列置換部１８に出力する。

【００６５】ここで日本語文書データの文字列を単語単位に文字列符号に変換するための辞書格納部１５を説明する。

【００６６】図9は、（株）日本電子化辞書研究所（ＥＤＲ）が研究成果として発表した、日本語を構成する形態素の品詞に関する集計結果である。この集計結果を見ると、単語数に対応する形態素数は１３６，４８６個であり、この単語の数を２進数で表現すると、最大表現数が２６２，１４３個となる１７ビットの符号で表すことができる。

【００６７】これに対し、新世代コンピュータ技術開発機構（ＩＣＯＴ）で作成した約１３万語の単語を有する日本語辞書から単語を構成する文字数を検出して分布を求めた結果、全登録単語の１／２以上の７万語が２文字で構成されており、平均文字数は２．８文字となっている。この平均文字数２．８文字をビット数で表すと、

２．８文字×２バイト＝５．６バイト×８ビット＝４４．８ビット

となる。

【００６８】そこで本発明にあっては、図9の１３６，４８６個の単語を表現する１７ビットの文字列符号を予め割り当て、入力した日本語データの文字列を単語単位に１７ビットの文字列符号に変換する符号化を行うことで、実質的にデータ量を半分以下に圧縮することができる。

【００６９】図１０は、図２の辞書格納部１５の辞書構造の実施形態である。図２の辞書格納部１５に格納された辞書は、先頭文字格納部４０と従属文字列格納部４２の２階層構造を備える。先頭文字格納部４０は、日本語文字「あ，い，う，え，お・・・」の文字コードをインデックスとしており、日本語の文字コードは２バイトデータであることから、文字コード４４としては、１６進数で「０ｘ０００」から「０ｘＦＦＦ」の１３１，０７２種類の格納位置が割り当てられる。

【００７０】この文字コード４４は、図２の文字列比較部１６で読み込んだ先頭文字を使用して、対応する文字コードの位置にアクセスする。文字コード２４に続いては先頭アドレス４６が格納される。先頭アドレス４６は、例えば文字コード４４の先頭文字「あ」を例にとると、先頭文字「あ」に続く従属文字列を格納した従属文字列格納部４２の先頭アドレス「Ａ１」を指定している。続いて従属文字列の個数４８が設けられる。例えば先頭文字「あ」にあっては、従属文字列個数４８としてＮ１＝４個が格納されている。

【００７１】従属文字列格納部４２は、先頭文字格納部４０の先頭文字の文字コード４４に対応して格納された先頭アドレス４６で先頭位置が指定され、この先頭位置から従属文字列格納部４８で指定された個数の格納位置に従属文字列が格納されている。例えば先頭文字「あ」に対応した先頭アドレス４６のアドレスＡ１から従属文字列個数４８のＮ１＝４個となる４つの格納位置が、対象とする従属文字列格納領域として指定される。

【００７２】この従属文字列格納部４２は、先頭から従属文字列の長さ５０、従属文字列５２、及び１７ビット表現される文字列コード（文字列符号）５４が格納されている。例えば先頭アドレスＡ１には、長さＬ１で従属文字列「い」と、その文字列コードが格納されている。次の格納位置には長さＬ２の従属文字列「う」がその文字列コードと共に格納されている。

【００７３】３番目の領域には長さＬ３の従属文字列「お」が文字列コードと共に格納されている。４番目の格納領域には長さＬ４で従属文字列が存在しないことを示す符号「ＮＵＬＬ」が格納され、存在しないことを示す文字列コードが格納されている。即ち、この４番目の格納領域は先頭１文字だけの文字列コードの登録を表している。

【００７４】ここで図１０の従属文字列格納部４２に文字列コード３４は、単語個数に基づき１番から１３６，４８６番まで予め１７ビットの文字列コードが割り当てられており、図１０のように格納した場合の文字列コード（文字列符号）Ｋと位置アドレスＸとの関係は、次式で表すことができる。

$$K=(N\cdot X-A1)/M \quad (1)$$

但し、Ｘ　：従属文字列格納部２２の位置アドレス

Ｎ　：一致検出された従属文字列の番号（１，２，３，・・Ｎ）

Ａ１：従属文字列格納部の開始アドレス

Ｍ　：従属文字列格納部の格納バイト長

ここで、従属文字列格納部４２の格納バイト長Ｍは、従属文字列の長さ５０、従属文字列５２、及び文字列コード５４の合計長であることから、例えば次式で表すこと

ができる。

格納バイト長M＝　長さ＋文字コード列＋文字列コード　（２）
＝3bit＋　96bit　＋　17bit
＝116bit
＝15byte

ここでは格納可能な従属文字列を最大６文字とすることで、従属文字列５２に９６ビットを割り当てた場合を例にとっている。もちろん、従属文字列の文字数は平均で２．８文字であることから、３文字（４８ビット）以上とすれば十分な圧縮効果が得られる。この場合には、従属文字列格納部の１つの格納領域の格納バイト長MはM

X＝M・K＋A1

但し、K　：文字列コード
A１：従属文字列格納部の開始アドレス
M　：復元側の従属文字列格納部の格納バイト長
この（３）式にあっては、復元側で使用する辞書における従属文字列格納部４２の開始アドレスA１、即ちオフセットと、従属文字列格納部４２の格納バイト長Mが定数として定まっていることから、復元しようとする文字列コードKを（３）式に代入することで、一義的に復元しようとする文字列を格納した辞書位置（位置アドレス）Xを算出することができる。

【００７６】図１１は図１０の辞書構造を持った辞書格納部１５による図２の文字列比較部１６による符号化処理のフローチャートである。

【００７７】まずステップS１で、文字列比較部１６に読み込んだ文字列の先頭文字の位置Pにポインタを移動し、ステップS２で先頭文字位置Pの文字コードが示す図１０の文字コード４４に対応した先頭文字格納部４０のテーブルを参照する。この先頭文字格納部４０のテーブル参照により、ステップS３で従属文字列格納部４２の先頭アドレス４６と従属文字列の個数４８を取得する。

【００７８】続いてステップS４で、従属文字列格納部４２の先頭アドレスの先頭データから従属文字列の長さ５０の長さデータLを取得する。次にステップS５で、先頭文字位置Pから従属文字列の長さデータLに基づくL文字を抽出し、抽出したL文字を従属文字列格納部４２の従属文字列５２の登録文字列と比較して、一致するか否か判断する。

【００７９】登録した従属文字列と一致すれば、ステップS３に進み、次の文字列コード５４を読み出し、一致検出した文字列に文字列比較部１６で割り当てて出力し、次のステップS９で先頭文字位置Pを従属文字列の文字数Lだけ移動した位置Pにポインタを更新する。そしてステップS１２で非圧縮データの処理が済んでいなければ、再びステップS２に戻り、更新した先頭文字位置Pについて同様な処理を繰り返す。

【００８０】一方、ステップS５で従属文字列格納部４２の登録従属文字列と一致しなかった場合には、従属文字列数Nが末了か否かチェックし、末了であればステップS７に戻り、従属文字列格納部４２の先頭アドレスの次の格納領域から従属文字列の長さデータLを取得し、ステップS５で再度、先頭文字位置PからL文字の従属文字列を抽出して従属文字列格納部４２の登録従属文字列と一致するか否か比較する。

【００８１】ステップS５～S７の繰返しによる登録個数Nの従属文字列の全てについて比較処理を行っても一致しなかった場合には、ステップS６で従属文字列の個数Nの終了を判別して、ステップS１０に進み、先頭文字１文字を表す未登録コードを送出する。そしてステップS１１で先頭文字位置Pを文字数L＝１文字だけ移動した次の位置にポインタを更新し、ステップS１２からステップS２に戻って、次の先頭文字位置Pからの処理を繰り返す。

【００８２】図１２は、図２のデータ圧縮装置から出力された符号処理部２６及びタグ情報ストリーム２８で構成される符号ストリームから文字列ストリームを復元するためのデータ復元装置の第１実施形態のブロック図である。

【００８３】このデータ復元装置は、タグ情報分離部６０、タグ情報格納部６２、文字列復元部６４で構成される。文字列復元部６４は、符号列比較部６６、辞書格納部７０及び文字列置換部６８を備える。

【００８４】タグ情報分離部６０は図２のデータ圧縮装置側から送られてきた符号ストリーム５８を入力し、タグ情報と符号データとに分離し、タグ情報はタグ情報格納部６２に格納し、符号データは符号ストリーム５８として文字列復元部６４に出力する。

【００８５】文字列復元部６４は符号列比較部６６で辞書格納部７０を用いて符号データから文字列及びタグ符号を復元した後に、文字列置換部６８においてタグ符号をタグ情報格納部６２に格納しているタグ情報に置き換えて、復元した文字列ストリーム７０を出力する。

【００８６】図１３は、図１２のデータ復元装置の復元処理のフローチャートである。まずステップS１で、タグ情報分離部６０が入力文書に対応した符号ストリーム５６からタグ情報を分離してタグ情報格納部６２に格納

する。次にステップS2で、タグ情報が分離された符号ストリーム58の中の符号列を辞書格納部65内の登録番号と比較照合し、一致する登録番号で格納している文字または文字列に変換する。

【0087】続いてステップS3で、復元された文字列に含まれているタグ符号をタグ情報格納部62に格納しているタグ情報の格納順に順次置換し、復元した文字列ストリーム70として出力する。これらのステップS1～S3の処理を、ステップS4で符号ストリーム56の入力が終了するまで繰り返す。

【0088】図12の文字列復元部64に設けた符号列比較部66は、辞書格納部65の参照により、図3のデータ圧縮装置で符号化された符号化ストリームから元の文字列を復元する。

【0089】図14は、図12の文字列辞書格納部70の辞書構造である。この文字列辞書格納部70にあっては、先頭文字72、従属文字列長さ74及び従属文字列76を、図10の辞書構造に示した従属文字列格納部42の17ビットの文字列コード54の順番に格納している。このため符号化比較部40にあっては、復元に使用する文字列辞書格納部42の格納バイト長Mが

$$\text{格納バイト長M} = \text{先頭文字} + \text{長さ} + \text{文字コード列} \quad (6)$$
$$= 16\text{bit} + 3\text{bit} + 96\text{bit}$$
$$= 115\text{ビット}$$
$$= 15\text{byte}$$

から判明しているため、次式から文字列コードKに対応した位置アドレスXを算出することができる。

$$X = M \cdot K + A1 \quad (7)$$

但し、K　：文字列コード

A1：文字列格納位置の開始アドレス

M　：格納バイト長

このようにして分離した文字列コードKから辞書格納位置を示す位置アドレスXを求めて参照することで、対応する先頭文字及び従属文字列を組み合わせた文字列を復元することができる。

【0090】このような図2のデータ圧縮装置及び図12のデータ復元装置により、図27に示したSGML日本語文書ファイルの文字列ストリームは、図6のようなタグ情報と図5のようなタグ情報をタグ符号に置き換えた文字列ストリームに分離され、この実施形態にあっては、タグ符号に置換済みの文字列ストリームを符号化することで文書ファイルの本文に相当する部分を高い圧縮率の圧縮ファイルに変換できる。

【0091】また図6のように分離されたタグ情報について、キーワードを使用して検索し、キーワードに一致するタグ情報が得られたならばタグ情報の出現位置が何番目かを検出し、これによって図5のタグ符号置換済みの本文の文書ファイルに含まれているタグ符号の出現位置を検索することで、タグ情報の検索結果に対応した文書位置の特定による読出し等が容易にできる。

【0092】図15は、本発明のデータ圧縮装置の第2実施形態であり、この実施形態にあっては、図2の第1実施形態に加えてタグ情報格納部78と符号格納部80を設けたことを特徴とする。

【0093】タグ情報格納部78にはタグ情報分離部10により文字列20から分離されたタグ情報が格納される。これによってタグ情報格納部78には、例えば図6のようなタグ情報ファイル36が格納される。また符号格納部80は文字列符号化部14に設けられており、タグ符号置換部12により分離したタグ情報にタグ情報を挿入したタグ置換済み文字列ストリーム22につき、図11の符号化処理により生成された符号データが格納される。

【0094】このようなタグ情報格納部78及び符号格納部80に加え、出力段に符号切替部82が設けられる。符号切替部82は、タグ情報格納部78に格納されたタグ情報と符号格納部80に格納された符号データを、例えば順番に選択して符号列ストリーム84として出力する。

【0095】図16は、図15のデータ圧縮装置の圧縮処理のフローチャートである。この圧縮処理は、ステップS1で、入力文書の文字列ストリーム20からタグ情報分離部10でタグ情報を分離し、タグ情報格納部78に格納する。次にステップS2で、文字列ストリーム20の中のタグのあった位置にタグ符号置換部12によって識別用のタグ符号を挿入する。

【0096】次にステップS3で、タグ符号の置換が済んだ文字列ストリーム22の文字列を文字列符号化部14の文字列比較部16に入力し、辞書格納部18内の辞書構造の対応する登録番号に変換する。このようなステップS1～S3の処理を、ステップS4で文字列ストリームの入力が終了するまで繰り返す。

【0097】文字列ストリームの入力が終了するとステップS5に進み、分離したタグ情報とタグ符号に変換して符号化した符号ストリームを、例えばタグ情報格納部78と符号格納部80から順番に読み出して符号ストリーム84として出力する。図15のデータ圧縮装置から出力された符号ストリーム84は、図12に示したデータ復元装置に入力することで文字列ストリームを復元することができる。

【0098】図17は、本発明のデータ圧縮装置の第3実施形態であり、この実施形態にあっては文字列ストリームから分離したタグ情報を圧縮するようにしたことを特徴とする。

【００９９】図１７において、このデータ圧縮装置は、図１５の第２実施形態におけるタグ情報分離部１０とタグ情報格納部７８の間に。新たにタグ情報圧縮部８６を設けている。タグ情報圧縮部８６は、タグ情報分離部１０において入力した文字列ストリーム２０から分離したタグ情報を圧縮対象の文字列ストリームとして圧縮してタグ情報格納部７８に格納する。

【０１００】タグ情報圧縮部８６による圧縮処理は、タグ情報にはタグと日本語文字列が含まれ、これらを一括して圧縮することから、ＬＺ７７、ＬＺ７８、算術符号化等の圧縮アルゴリズムを使用する。

【０１０１】タグ情報分離部１０、タク符号置換部１２、文字列符号化部１４は、図１５の第２実施形態と同じである。

【０１０２】図１８は、図１７のデータ圧縮装置による圧縮処理の説明図である。ＳＧＭＬ日本語文書ファイル３５の内容となる文字列ストリーム２０は、タグ情報分離部１０によってタグ情報ファイル３６の内容となるタグ情報に分離される。このタグ情報はタグ情報圧縮部８６により圧縮した後、タグ情報格納部７８の格納を介して出力する。

【０１０３】またＳＧＭＬ日本語文書ファイル３５の内容となる文字列ストリーム２０から分離したタグ情報の位置には、タグ符号置換部１２によって固定タグ符号または出現順序を示す順序タグ符号が挿入配置され、タグ置換済み日本語文書ファイル３８の内容となる文字列ストリーム２２が文字列符号化部１４に出力され、文字列符号化により圧縮された符号データが符号格納部８０による格納を介して出力される。

【０１０４】図１９は、図１７のデータ圧縮装置から出力された符号ストリーム９０から文字列ストリームを復元する本発明のデータ復元装置の第２実施形態である。このデータ復元装置は、図１２の第１実施形態に更に圧縮タグ格納部６２とタグ情報復元部９２を設けている。

【０１０５】タグ情報分離部６０は、入力する符号ストリーム９０に含まれる圧縮タグ情報を分離して圧縮タグ格納部６２に格納する。圧縮タグ格納部６２に格納された圧縮タグ情報はタグ情報復元部９２により復元され、タグ情報格納部６２に格納される。タグ情報復元部９２はデータ圧縮側のＬＺ７７，ＬＺ７８，算術復号化に対応した復元アルゴリズムを実行する。それ以外の構成は図１５と同じになる。

【０１０６】図２０は、本発明のデータ圧縮装置の第４実施形態であり、分離したタグ情報の中の日本語文字列を符号化により圧縮し、更に、分離したタグ情報に本文中の置き換えを行ったタグ符号の位置を示す位置指定情報を付加するようにしたことを特徴とする。

【０１０７】図２０において、タグ情報分離部１０、タグ符号置換部１２、文字列比較部１６、辞書格納部１８を備えた文字列符号化部１４、タグ情報格納部７８及び符号出力部８２は、図１５の第２実施形態と同じである。これに加えて図２０の第４実施形態にあっては、新たにタグ文字列比較部９４、タグ辞書格納部９６及び符号量計測部９８を設けている。

【０１０８】タグ文字列比較部９４とタグ辞書格納部９６は、タグ情報分離部１０で分離したタグ情報に含まれる日本語文字列ストリームを文字列符号化部１４と同様な符号化アルゴリズムで符号化してタグ情報を圧縮する。このため、タグ情報格納部９６の辞書構成は図１０と同じであり、先頭文字及び従属文字としてタグ情報に使用する日本語文字列が使用されている。またタグ文字列の符号化処理は図１１のフローチャートに従って行う。

【０１０９】一方、図２０のデータ圧縮装置に設けた符号量計測部９８は、文字列符号化部１４による本文の文字列ストリーム２２、即ちタグ符号の置換が済んだ文字列ストリーム２２を対象とした符号化による符号データについて、文字列ストリームの先頭から置換済みの各タグ符号までの符号量を計測し、この各タグ符号までの符号量の計測結果を、タグ情報格納部７８に格納する文字列ストリームから分離した各タグ情報のそれぞれに符号位置情報として付加して格納する。

【０１１０】符号量計測部９８によるタグ符号で置換されたタグ情報の位置を示す位置指定情報としては、文字列ストリームの先頭からの符号量以外に、文字列ストリームの中に特定のタグ情報からの後続する各タグ情報までの符号データの符号量としてもよい。

【０１１１】図２１は、図２０の第４実施形態における圧縮処理の説明図である。ＳＧＭＬ日本語文書ファイル３５の内容となる文字列ストリームを入力して、タグ情報の分離によるタグ情報ファイル３６の生成及びタグ情報をタグ符号に置換したタグ置換済み日本語文書ファイル３７の生成は、図１５の第２実施形態と同じである。

【０１１２】これに加えて、分離されたタグ情報ファイル３６のタグ情報に含まれている日本語文字列であるタグ文字列を、タグ辞書格納部９６を用いて符号化して圧縮することで出力している。

【０１１３】図２２は、タグ情報格納部７８に格納されたタグ情報ファイルの具体例であり、図２７に示したＳＧＭＬ日本語文書ファイルから分離したタグ情報を例にとっている。このタグ情報ファイル３６には、左側のインデックス０１～１３に対応した各タグに対応して、右側に図２１のタグ置換済み日本語ファイル３７の文字列データの符号データの先頭からの符号量（バイト量）ＤＬ１～ＤＬ１３が位置指定情報１０６としてそれぞれ格納されている。

【０１１４】図２３は、図２０の第４実施形態による圧縮処理のフローチャートである。まずステップＳ１～Ｓ４は図８と同じであり、タグ情報分離部１０で文字列ストリーム２０から分離したタグ情報をタグ情報格納部７

８に格納し、またタグ符号置換部１２により分離したタグ情報の位置にタグ符号２４を挿入配置した文字列ストリーム２２を文字列符号化部１４で符号化して符号データを符号格納部８０に格納する。

【０１１５】次のステップＳ４にあっては、符号量計測部９８が文字列符号化部１４で置換済みのタグ符号を符号化する際に、例えば文字列ストリームの先頭からの符号量ＤＬを計測し、既にタグ情報格納部７８に格納されているタグ情報に図２２の位置指定情報１０６として計測した符号量ＤＬを格納する。

【０１１６】このようなステップＳ１～Ｓ４の処理を、ステップＳ５で文字列ストリームの入力が終了するまで繰り返す。文字列ストリーム２０の入力が終了すると、ステップＳ６で、タグ情報格納部７８に分離したタグ情報中の文字列をタグ辞書格納部９６内の辞書の対応するブロック番号に変換して符号データとする符号化処理をタグ文字列比較部９４で行い、タグ情報格納部７８に格納する。その結果、タグ情報格納部７８の格納内容は図２２の圧縮タグ情報ファイル３６のようになる。

【０１１７】最後にステップＳ７で、タグ情報格納部７８に分離して符号化した符号量付きのタグ情報と符号化部８０に格納した符号データを符号切替部８２より例えば順番に選択出力し、符号ストリーム１００として外部に供給する。

【０１１８】ここで、図２３の圧縮処理にあっては、ステップＳ１～Ｓ４のタグ情報の分離と置換、更には圧縮した符号量の計測処理とその後の分離したタグ情報の符号化処理を、時間的に分けて処理しているが、両者を並行して処理するようにしてもよいことはもちろんである。

【０１１９】図２４は、図２０のデータ圧縮装置から出力された符号ストリーム１００から文字列ストリームを復元する本発明のデータ復元装置の第３実施形態である。

【０１２０】図２４において、タグ情報分離部６０、圧縮タグ格納部９２、タグ情報格納部６２、文字列復元部６４は図１９の第２実施形態と同じであり、これに加えて図２４の第３実施形態にあっては、タグ文字列復元部１０２とタグ復元辞書格納部１０４を新たに設けている。

【０１２１】タグ復元辞書格納部１０４は、図１３の辞書構造と同じものが使用され、格納している文字がタグに使用している日本語文字列となっている。タグ情報分離部６０は、図２０のデータ圧縮装置側から供給される符号ストリーム１００から、図２２の圧縮タグ情報ファイル３６の内容に示すようなタグ情報ストリームを分離し、圧縮タグ格納部９２に格納する。

【０１２２】圧縮タグ格納部９２に格納された圧縮タグ情報は、タグ文字列復元部１０２によるタグ復元辞書格納部１０４のタグ文字列の符号による辞書番号の参照で対応する日本語文字列に復元され、復元した日本語文字列を含むタグ情報をタグ情報格納部６２に格納する。

【０１２３】一方、タグ情報分離部６０は、圧縮タグ情報ストリームに続いて送られてくる文書本文の符号ストリームを文字列復元部６４に供給し、符号列比較部６６で取り出した符号による辞書格納部６５の辞書番号の参照で対応する文字または文字列を復元し、符号置換部６８に出力する。

【０１２４】符号置換部６８は、復元した文字列の中のタグ符号を認識し、その出現順に従ってタグ情報格納部６２に格納している復元済みのタグ情報を格納順に取り出し、タグ符号と置換し、復元した文字列ストリームを出力する。

【０１２５】ここで、圧縮タグ格納部９２には、符号ストリーム１００から分離した圧縮タグ情報ストリームの入力が終了した時点で、図２２のように圧縮タグ情報ファイル３６が格納されている。そこで圧縮タグ情報ファイル３６について、特定のタグをキーワードとして検索を行い、一致するタグが得られたならば、これに対応する位置指定情報としての符号量ＤＬを読み出し、図２０のデータ圧縮装置に対し検索した符号量ＤＬの位置からの符号データの転送要求ができる。これによって、データ復元側からデータ圧縮側に必要とするＳＧＭＬ日本語文書の部分的な圧縮本文データの転送による読込みが簡単にできる。

【０１２６】尚、本発明におけるデータ圧縮装置からデータ復元装置への伝送形態としては、インターネット等の通信回線でもよいし、光ディスクカートリッジや磁気ディスクカートリッジ等の書替可能な可搬媒体等の適宜の形態でよい。

【０１２７】また上記の実施形態は、タグ情報を分離し、分離したタグ情報の位置にタグ符号を置換した文字列ストリームの圧縮として、日本語固有の単語数に対応した固定長の文字列符号を割り当てた符号化を例にとっているが、これ以外のＬＺ７７，ＬＺ７８，算術符号化等の圧縮を行うようにしてもよいことはもちろんである。

【０１２８】更に、本発明は、上記の実施形態の数値による限定は受けない。更に本発明は、その目的と利点を損なわない範囲の適宜の変形を含む。

【０１２９】

【発明の効果】以上説明してきたように本発明によれば、タグを含むＳＧＭＬ等の構造化文書の文字列ストリームにつき、タグ情報と本文（文字列）とを分離し、少なくとも本文を符号化することで高い圧縮率を実現し、また分離したタグ情報を検索することで圧縮された符号データ中の特定のタグ位置の読出しや検索を高速で処理することができる。

【０１３０】即ち、分離したタグ情報の順番と符号データ中に置換したタグ符号の順番は１対１に対応してお

り、タグ情報について特定のタグ情報を検索することで、その順番から符号データ中のタグ符号の位置が特定でき、目標とする文書符号データの先頭位置に容易に到達することができる。

【０１３１】この結果、タグを含むＳＧＭＬ等の構造化文書について、高い圧縮率を保ちながら高速に圧縮及び復元を行うことができる。

【図面の簡単な説明】

【図１】本発明の原理説明図

【図２】本発明のデータ圧縮装置の第１実施形態のブロック図

【図３】図２のタグ情報分離部のブロック図

【図４】図２のデータ圧縮装置の処理手順の説明図

【図５】図４のタグをタグ符号に置換した本文ファイルの説明図

【図６】図４の文字列ストリームから分離したタグ情報ファイルの説明図

【図７】図４のタグを出現順序付きのタグ符号に置換した本文ファイルの説明図

【図８】図２のデータ圧縮装置の圧縮処理のフローチャート

【図９】日本語文書に対する研究結果の説明図

【図１０】図２の辞書格納部の辞書構造の説明図

【図１１】図１０の辞書構造を用いた図２の符号化処理のフローチャート

【図１２】図２のデータ圧縮装置からの符号ストリームを復元する本発明のデータ復元装置の第１実施形態のブロック図

【図１３】図１２のデータ復元装置の復元処理のフローチャート

【図１４】図１２の辞書格納部の辞書構造の説明図

【図１５】本発明のデータ圧縮装置の第２実施形態のブロック図

【図１６】図１５のデータ圧縮装置の圧縮処理のフローチャート

【図１７】本発明のデータ圧縮装置の第３実施形態のブロック図

【図１８】図１７のデータ圧縮装置の処理手順の説明図

【図１９】図１７のデータ圧縮装置からの符号ストリームを復元する本発明のデータ復元装置の第２実施形態のブロック図

【図２０】本発明のデータ圧縮装置の第４実施形態のブロック図

【図２１】図２０のデータ圧縮装置の処理手順の説明図

【図２２】図２１の符号量をタグに付加した図２０のデータ圧縮装置に格納されるタグ情報ファイルとタグ情報ストリームの説明図

【図２３】図２０のデータ圧縮処理のフローチャート

【図２４】図２０のデータ圧縮装置からの符号ストリームを復元する本発明のデータ復元装置の第３実施形態のブロック図

【図２５】ＳＧＭＬ文書の構造説明図

【図２６】ＳＧＭＬ文書の文書型定義ＤＴＤの具体例の説明図

【図２７】日本語文書を例にとったＳＧＭＬ文書ファイルの説明図

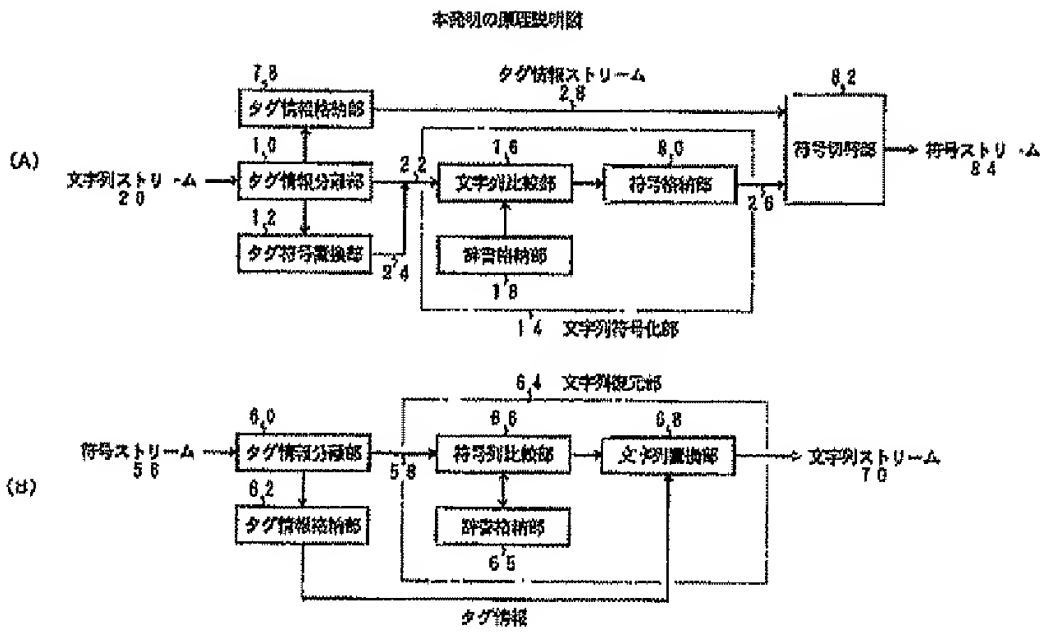
【図２８】ＳＧＭＬ文書ファイルを圧縮する基本的な符号化アルゴリズムのフローチャート

【図２９】圧縮していないタグ情報の部分と圧縮した本文部分が混在したＳＧＭＬ文書圧縮データファイルの説明図

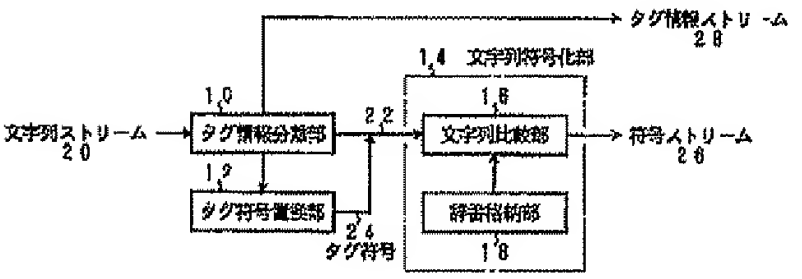
【符号の説明】

１０：タグ情報分離部
１２：タグ符号置換部
１４：文字列符号化部
１６：文字列比較部
１８：辞書格納部
２０：文字列ストリーム
２２：タグ置換済み文字列ストリーム
２４：タグ符号
２６，８４，９０，１０２：符号ストリーム
２８，１００：タグ情報ストリーム
３０：タグ比較部
３２：タグ識別規則格納部
３４：出力切替部
３５：ＳＧＭＬ日本語文書ファイル
３６：タグ情報ファイル
３８：タグ置換済み日本語文書ファイル
４０：先頭文字格納部
４２：従属文字列格納部
６０：タグ符号分離部
６２：タグ情報格納部
６４：文字列復元部
６６：符号列比較部
６８：文字列置換部
７０：辞書格納部
７８：タグ情報格納部
８６：タグ情報圧縮部
８８：圧縮タグ情報ストリーム
９２：圧縮タグ格納部
９４：タグ情報復元部
９５：符号量計測部
９６：タグ文字列比較部
９８：タグ辞書格納部
９８：符号量計測部
１０４：タグ文字列復元部
１０６：タグ復元辞書格納部

【図１】


本発明の原理説明図
(A)
文字列ストリーム 20
タグ情報分離部 10
タグ符号置換部 12
タグ情報格納部 18
タグ情報ストリーム 28
文字列比較部 16
符号格納部 30
辞書格納部 18
文字列符号化部 14
符号切替部 32
符号ストリーム 34
(B)
符号ストリーム 56
タグ情報分離部 60
タグ情報格納部 62
符号列比較部 66
辞書格納部 65
文字列置換部 68
文字列ストリーム 70
文字列復元部 64
タグ情報

【図２】


本発明のデータ圧縮装置の第１実施形態のブロック図
タグ情報ストリーム 28
文字列ストリーム 20
タグ情報分離部 10
タグ符号置換部 12
文字列符号化部 14
文字列比較部 16
辞書格納部 18
符号ストリーム 26
タグ符号 24

【図３】

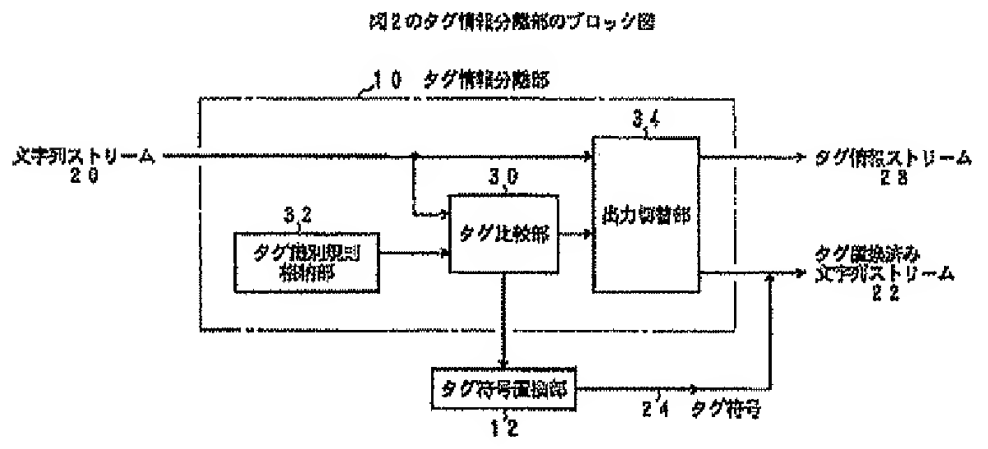


図２のタグ情報分離部のブロック図
10 タグ情報分離部
文字列ストリーム 20
タグ比較部 30
タグ識別規則格納部 32
出力切替部 34
タグ情報ストリーム 28
タグ置換済み文字列ストリーム 22
タグ符号置換部 12
タグ符号 24

【図１４】

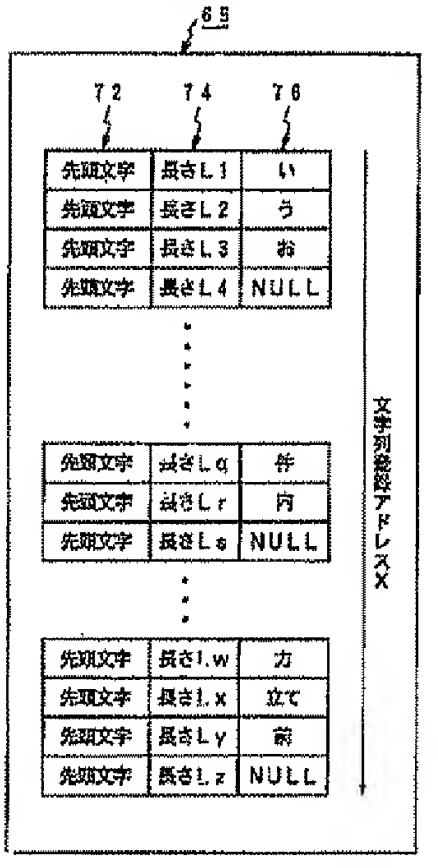
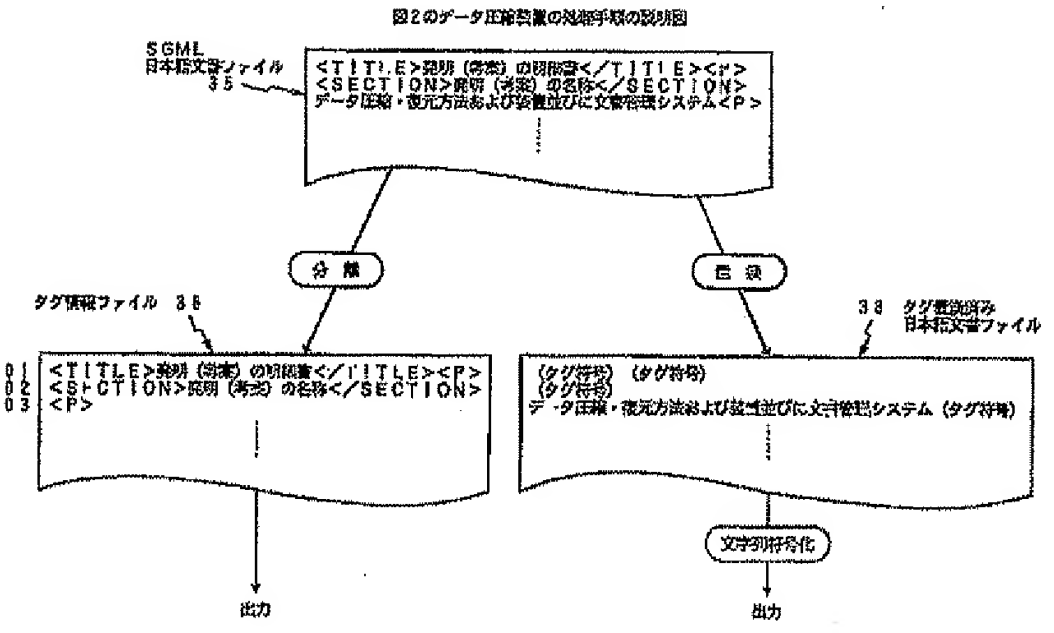


図１２の辞書格納部の辞書構造の説明図
65
72
74
76
先頭文字 長さL1 い
先頭文字 長さL2 う
先頭文字 長さL3 お
先頭文字 長さL4 NULL
先頭文字 長さL.α 件
先頭文字 長さL.r 内
先頭文字 長さL.s NULL
先頭文字 長さL.w 力
先頭文字 長さL.x 立て
先頭文字 長さL.y 前
先頭文字 長さL.z NULL
文字列格納アドレスX

【図４】



【図５】

図４のタグをタグ符号に置換した本文ファイルの説明図

３８

(タグ符号) (タグ符号)
(タグ符号)
データ圧縮・復元方法および装置並びに文書管理システム (タグ符号)
(タグ符号)
(タグ符号)
(タグ符号)
を有することを特徴としたデータ圧縮方法。(タグ符号)
(タグ符号)
(タグ符号)
(タグ符号) 近年、文字コード、ベクトル化情報、画像など様々な種類のデータがコンピュータで扱われるようになっており、扱われるデータ量も急速に増加してきている。大きのデータを扱うときは、データの中の冗長な部分を省いてデータ量を圧縮することで、記憶容量をへらしたり、速く伝送したりできるようになる。
(タグ符号) 一方、最近では計算機で扱う文書の形式を統一する動きがある。これまで、計算機あるいはアプリケーションによってばらばらであった文書の形式を異なる計算機でも使用できるようにするものである。(タグ符号) は１９９８年にＩＳＯ・・・

【図６】

図４の文字列ストリームから分離したタグ情報ファイルの説明図

３６

```
０１ <TITLE>発明（考案）の明細書</TITLE>
０２ <P>
０３ <SECTION>発明（考案）の名称</SECTION>
０４ <P>
０５ <SECTION>請求項目の範囲</SECTION>
０６ <SUBSECTION>入力データを符号化するデータ圧縮装置において、</SUBSECTION>
０７ <LIST>
       <ITEM>符号表を保持する手段と
       <ITEM>符号表を元に入力データを符号化する処理と
     </LIST>
０８ <P>
０９ <SECTION>発明の詳細な説明</SECTION>
１０ <SUBSECTION>産業上の利用分野</SUBSECTION>
１１ <PARAGRAPH>
１２ <PARAGRAPH>
１３ <TT>SGML (Standard Generalized Markup Language) </TT>
     ・・・
```

【図7】

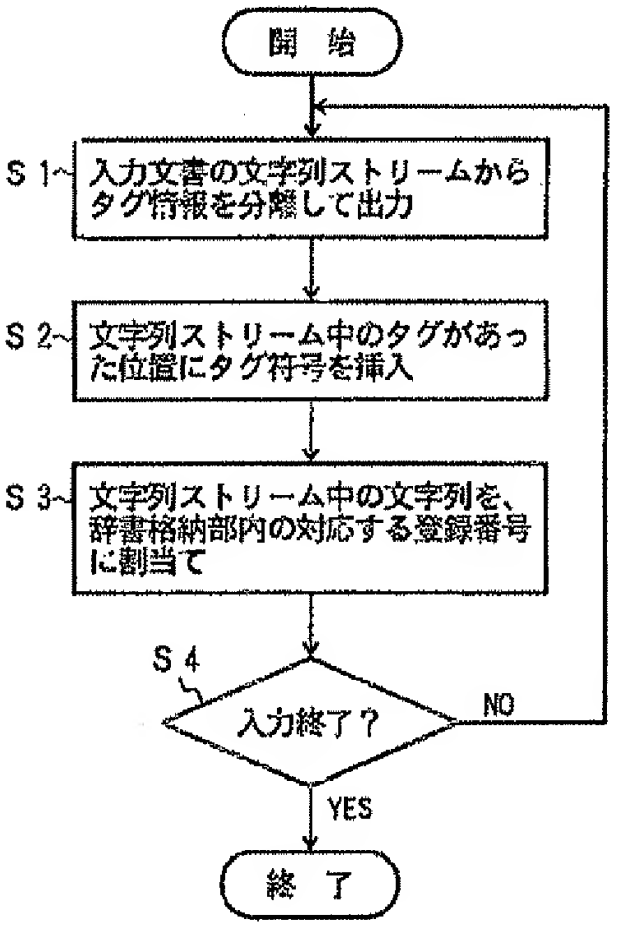
図4のタグを出現順序付きのタグ符号に置換した本文ファイルの説明図

38

(タグ符号1) (タグ符号2)
(タグ符号3)
データ圧縮・復元方法および装置並びに文書管理システム (タグ符号4)
(タグ符号5)
(タグ符号6)
(タグ符号7)
を有することを特徴としたデータ圧縮方法。(タグ符号8)
(タグ符号9)
(タグ符号10)
(タグ符号11) 近年、文字コード、ベクトル情報、画像など様々な種類のデータがコンピュータで扱われるようになっており、扱われるデータ量も急速に増加してきている。大量のデータを扱うときは、データの中の冗長な部分を省いてデータ量を圧縮することで、記憶容量をへらしたり、速く伝送したりできるようになる。
(タグ符号12) 一方、最近では計算機で扱う文書の形式を統一する動きがある。これまで、計算機あるいはアプリケーションによってばらばらであった文書の形式を異なる計算機でも使用できるようにするものである。(タグ符号13)は1986年にISO・・・

【図8】

図2のデータ圧縮装置の圧縮処理のフローチャート



【図25】

SGML文書の構造説明図

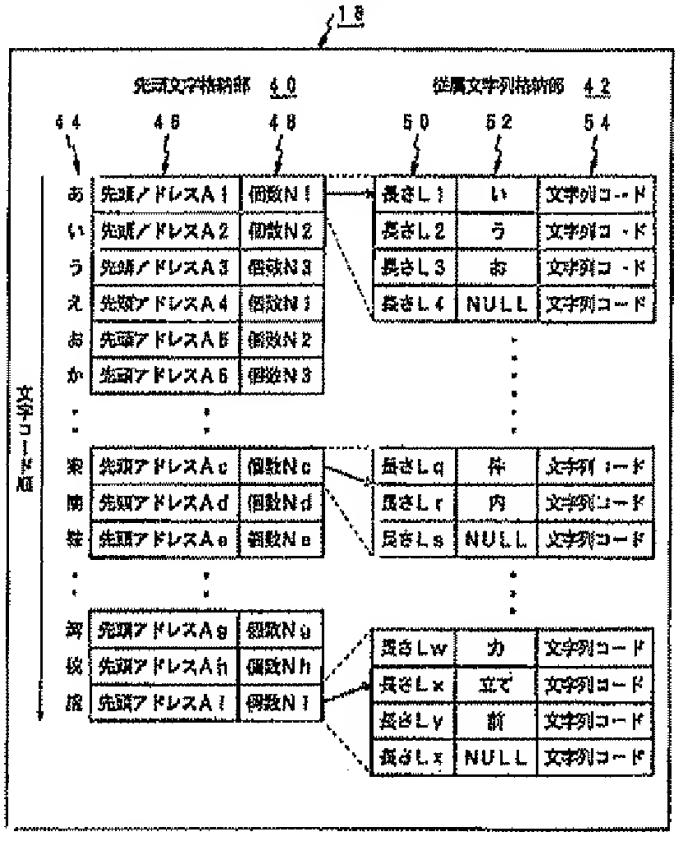
【図9】

日本語文書に対する研究結果の説明図

品詞類	形態素数(総数)	構成比(%)	形態素数(異なり数)	構成比(%)
名詞類	1,375,378	26.1	110,912	81.3
動詞類	622,125	11.8	14,538	10.7
形容詞類	58,742	1.1	1,204	0.9
形容動詞類	61,192	1.2	3,796	2.8
副詞類	74,332	1.4	2,934	2.1
連体詞類	40,271	0.8	247	0.2
接続詞類	23,582	0.4	247	0.2
感動詞類	21,063	0.4	318	0.2
接尾類	122,954	2.3	1,339	1.0
語尾類	631,304	12.0	155	0.1
助詞類	1,402,757	26.7	171	0.1
助動詞類	319,852	6.1	203	0.1
感動詞類	350	0.0	105	0.1
その他	508,333	9.7	226	0.2
総 計	5,282,221	100.0	136,486	100.0

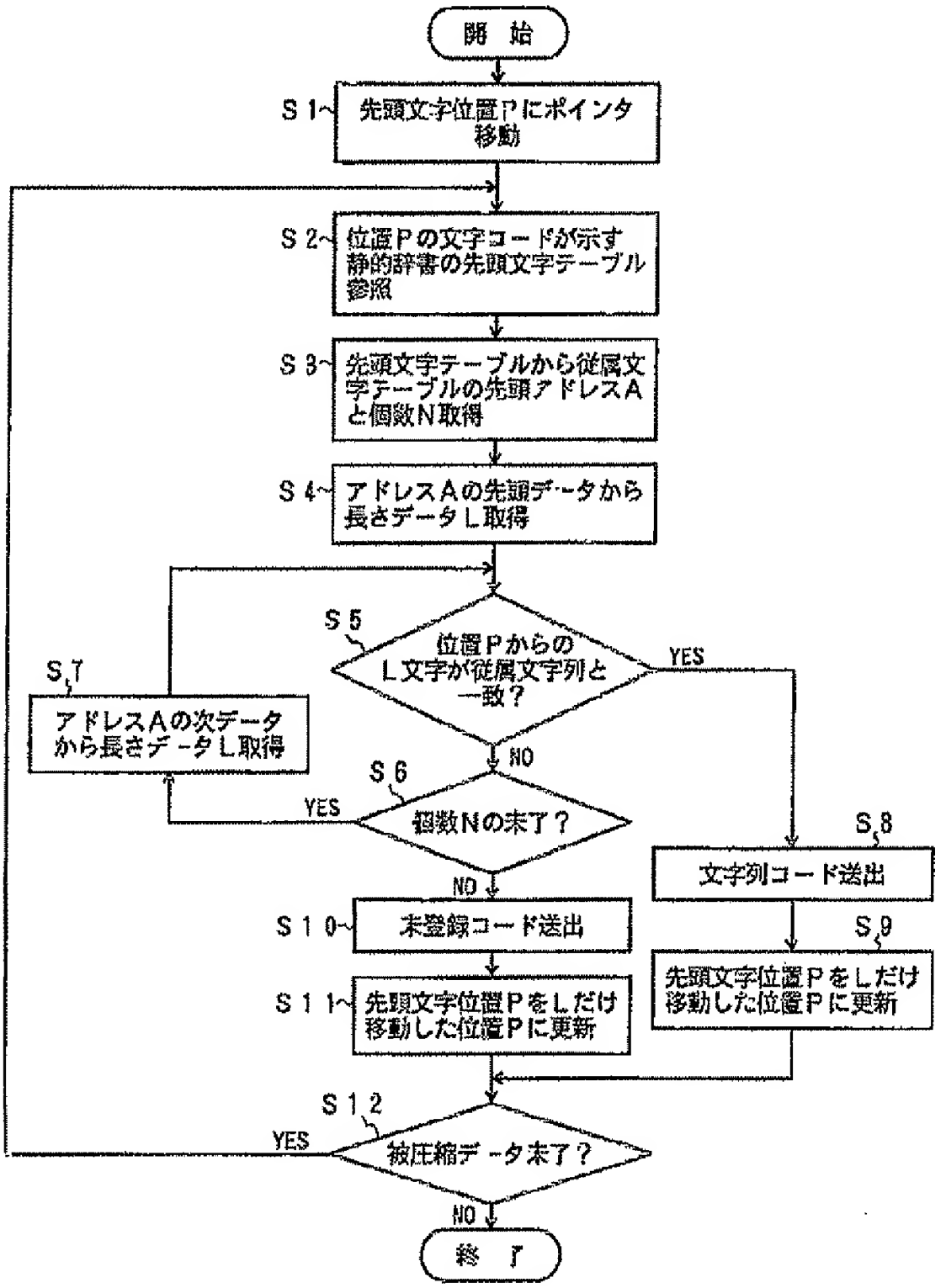
【図10】

図2の辞書格納部の辞書構造の説明図



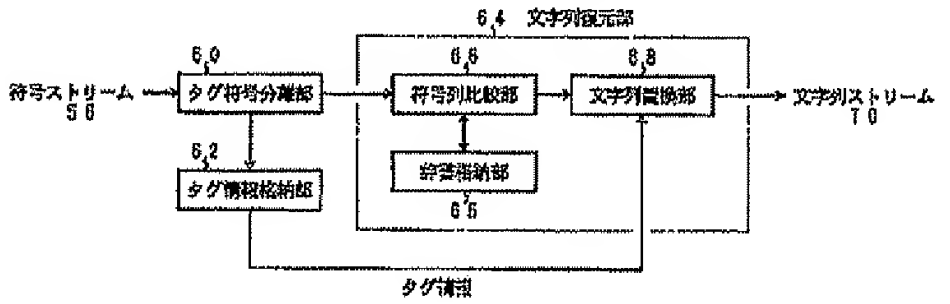
【図11】

図10の辞書構造を用いた図2の符号化処理のフローチャート



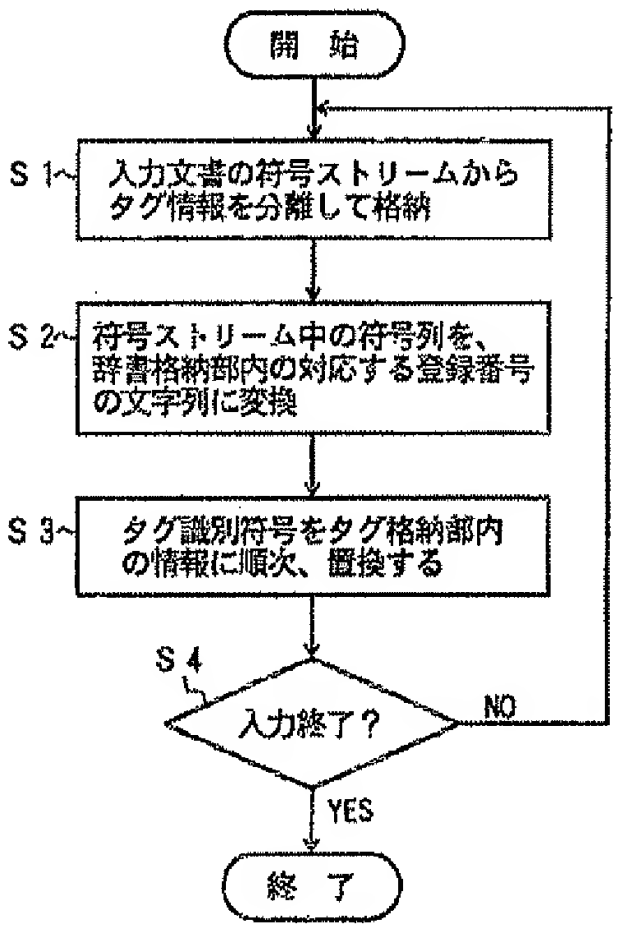
【図12】

図2のデータ圧縮装置からの符号ストリームを復元する本発明のデータ復元装置の第1実施形態のブロック図



【図13】

図12のデータ復元装置の復元処理のフローチャート



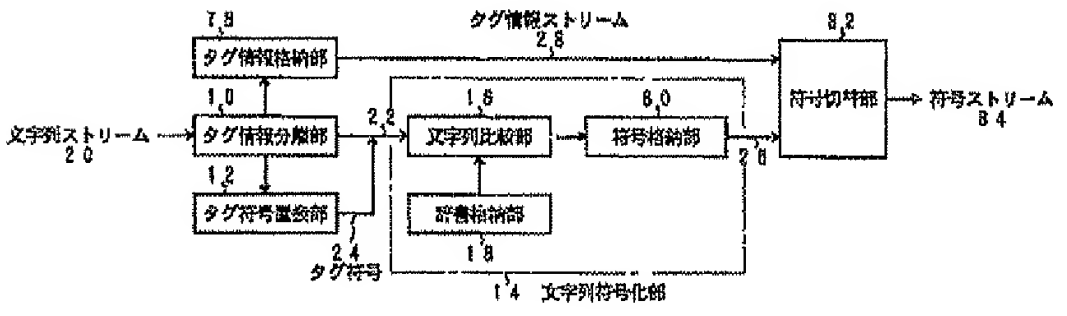
【図26】

SGML文書の文書型定義DTDの具体例の説明図

```
                    ********
<!--============Link Markup ============================-->
<!ENTITY % linkType "NAME">
<!ENTITY % linkExtraAttributes
        "REL %linkType;#IMPLIED
        REV %linkType;#IMPLIED
        URN CDATA#IMPLIED
        TITLE CDATA#IMPLIED
        METHODS NAMES#IMPLIED
        ">
<![% HTML.Recommended [
        <!ENTITY % A.content  "(%text)* "
        --<H1><a name="xxx">Heading</a></H1>
              is preferred to
          <a name="xxx"></H1>Heading</H1></a>
        -->
]]>
<!ENTITY % A.content  "(%heading | %text)* ">
<!ELEMENT A        --%A.content-(A)>
<!ATTLIST A
        HREF CDATA#IMPLIED
        NAME CDATA#IMPLIED
        %linkExtraAttributes;
        %SDAPR#F;"<Anchor:#AttList>"
<!--<A>              Anchor:source/destination of link   -->
<!--<A NAME="...">   Name of this anchor                 -->
<!--<A HREF="...">   Address of link destination        -->
<!--<A URN="...">    Permanent address of destination   -->
                    ********
```

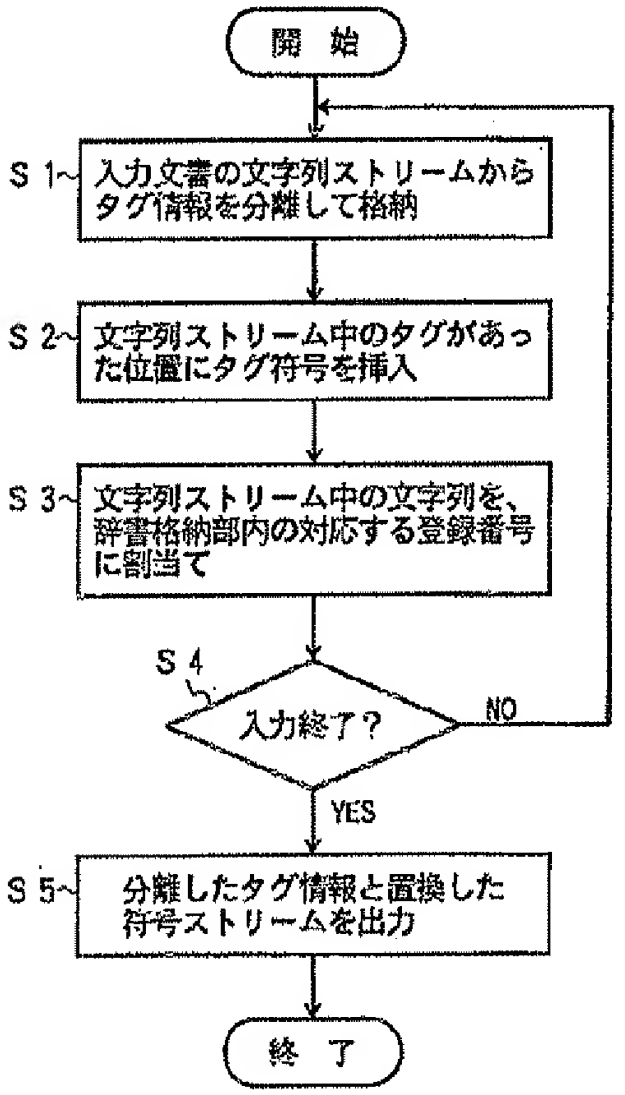
【図15】

本発明のデータ圧縮装置の第2実施形態のブロック図



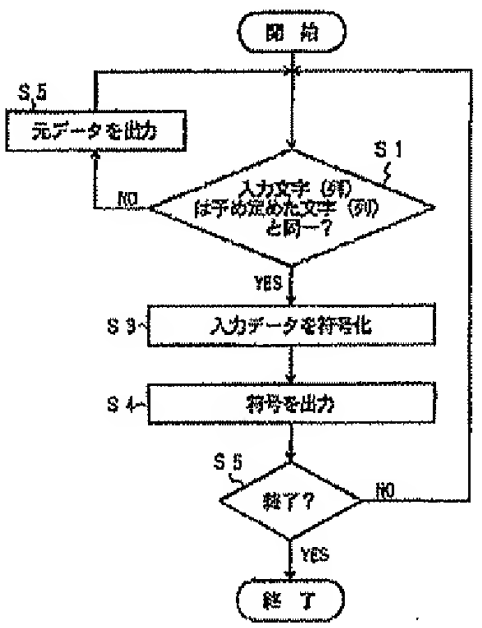
【図１６】

図１５のデータ圧縮装置の圧縮処理のフローチャート



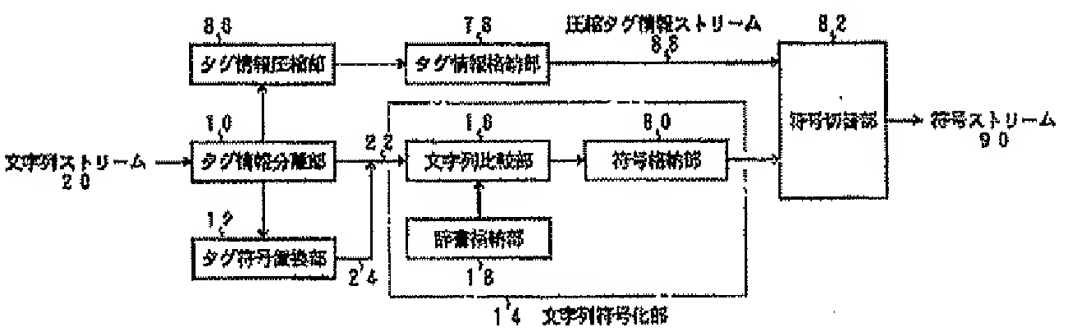
【図２８】

ＳＧＭＬ文書ファイルを圧縮する基本的な符号化アルゴリズムのフローチャート



【図１７】

本発明のデータ圧縮装置の第３実施形態のブロック図



【図18】

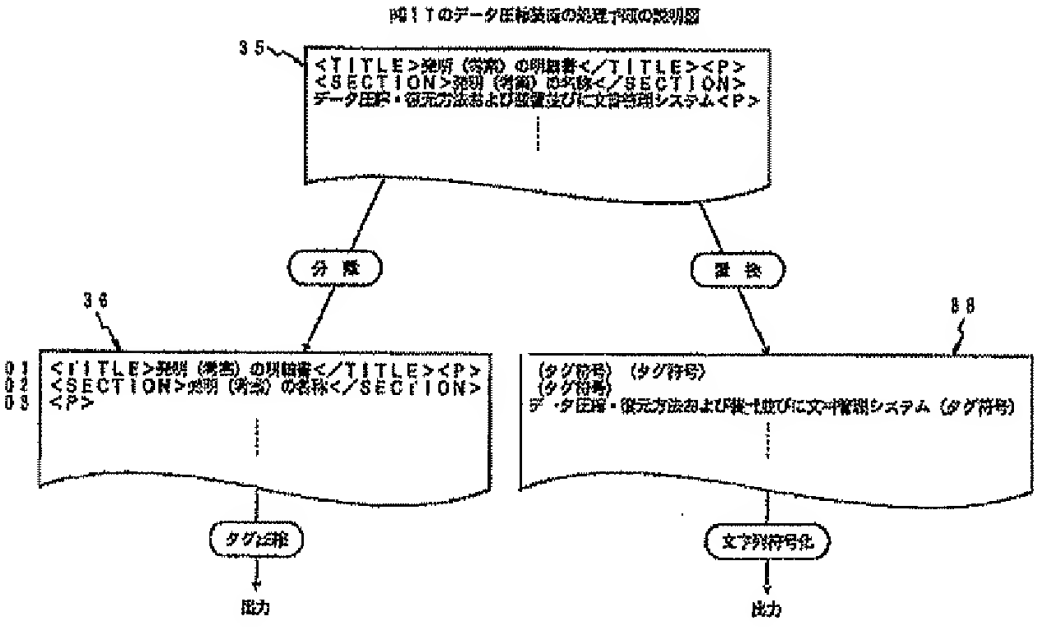
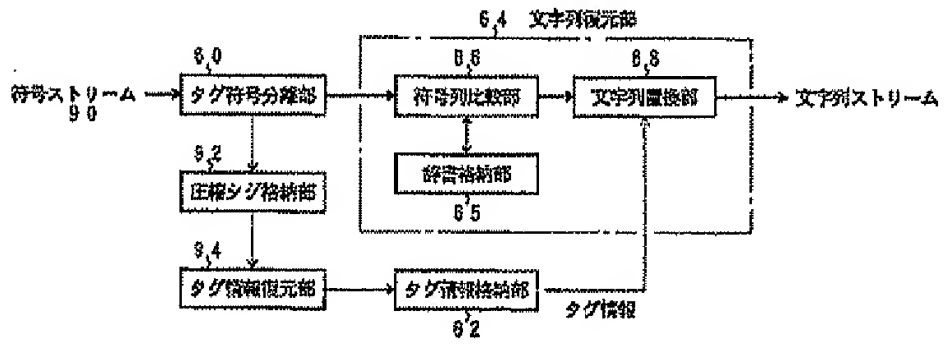
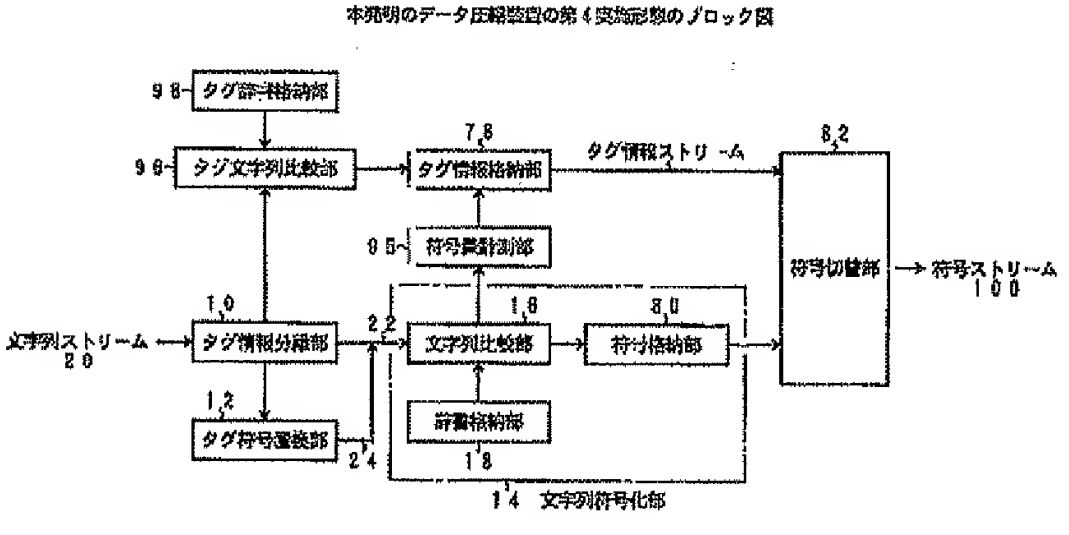


図11のデータ圧縮装置の処理手順の説明図
35
<TITLE>発明（考案）の明細書</TITLE><P>
<SECTION>発明（考案）の名称</SECTION>
データ圧縮・復元方法および装置並びに文書管理システム<P>
分 離
置 換
36
38
01 <TITLE>発明（考案）の明細書</TITLE><P>
02 <SECTION>発明（考案）の名称</SECTION>
03 <P>
（タグ符号）（タグ符号）
（タグ符号）
データ圧縮・復元方法および装置並びに文書管理システム（タグ符号）
タグ圧縮
文字列符号化
出力
出力

【図19】

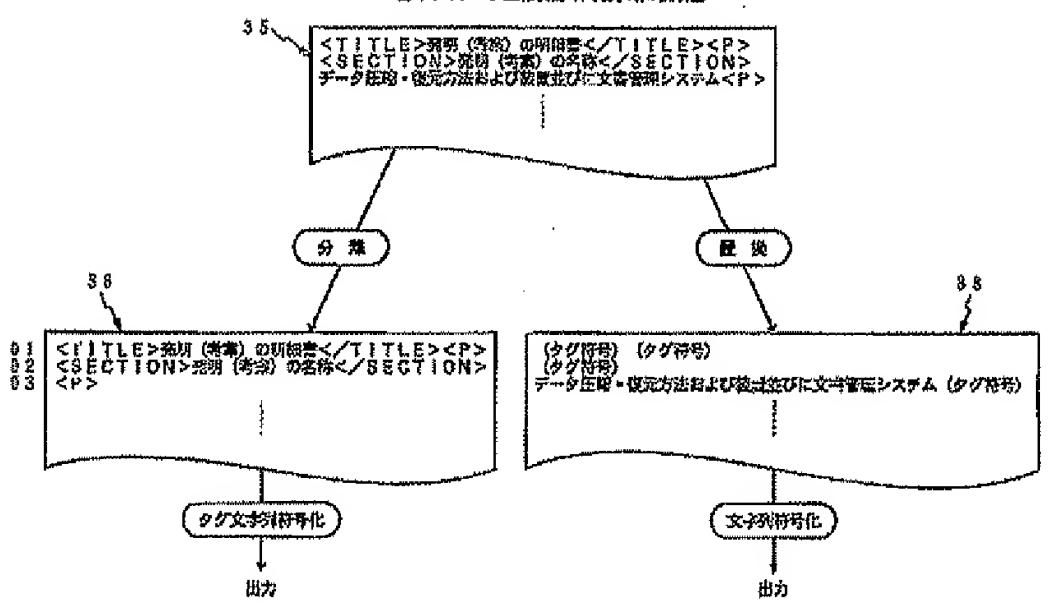

64 文字列復元部
60
符号ストリーム
90
タグ符号分離部
66
符号列比較部
68
文字列置換部
文字列ストリーム
62
圧縮タグ格納部
辞書格納部
65
64
タグ情報復元部
タグ情報格納部
62
タグ情報

【図20】


本発明のデータ圧縮装置の第4実施形態のブロック図
98
タグ辞書格納部
96
タグ文字列比較部
78
タグ情報格納部
タグ情報ストリーム
82
85
符号量計測部
符号切替部
符号ストリーム
100
10
タグ情報分離部
22
文字列比較部
18
80
符号格納部
文字列ストリーム
20
12
タグ符号置換部
24
辞書格納部
18
14 文字列符号化部

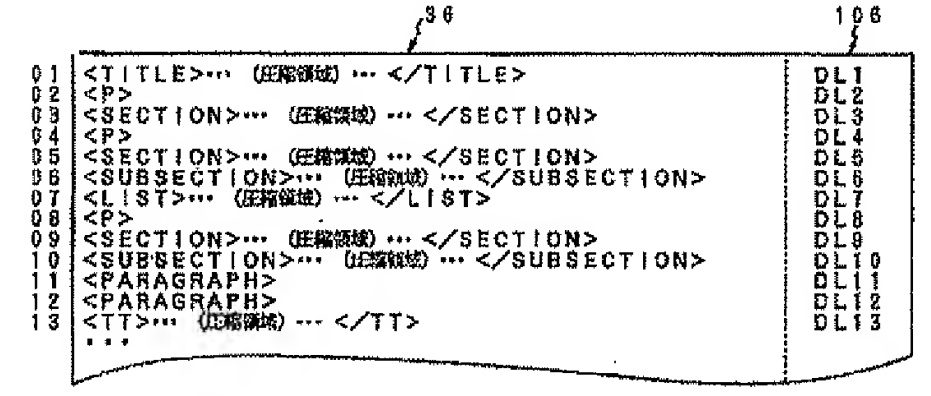
【図21】

図20のデータ圧縮装置の処理手順の説明図



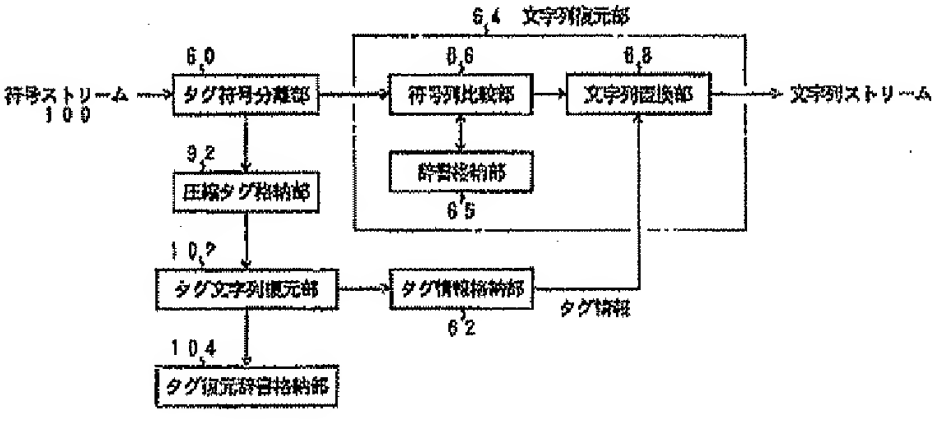
【図22】

図21の符号量をタグに付加した図20のデータ圧縮装置に格納されるタグ情報ファイルとタグ情報ストリームの説明図



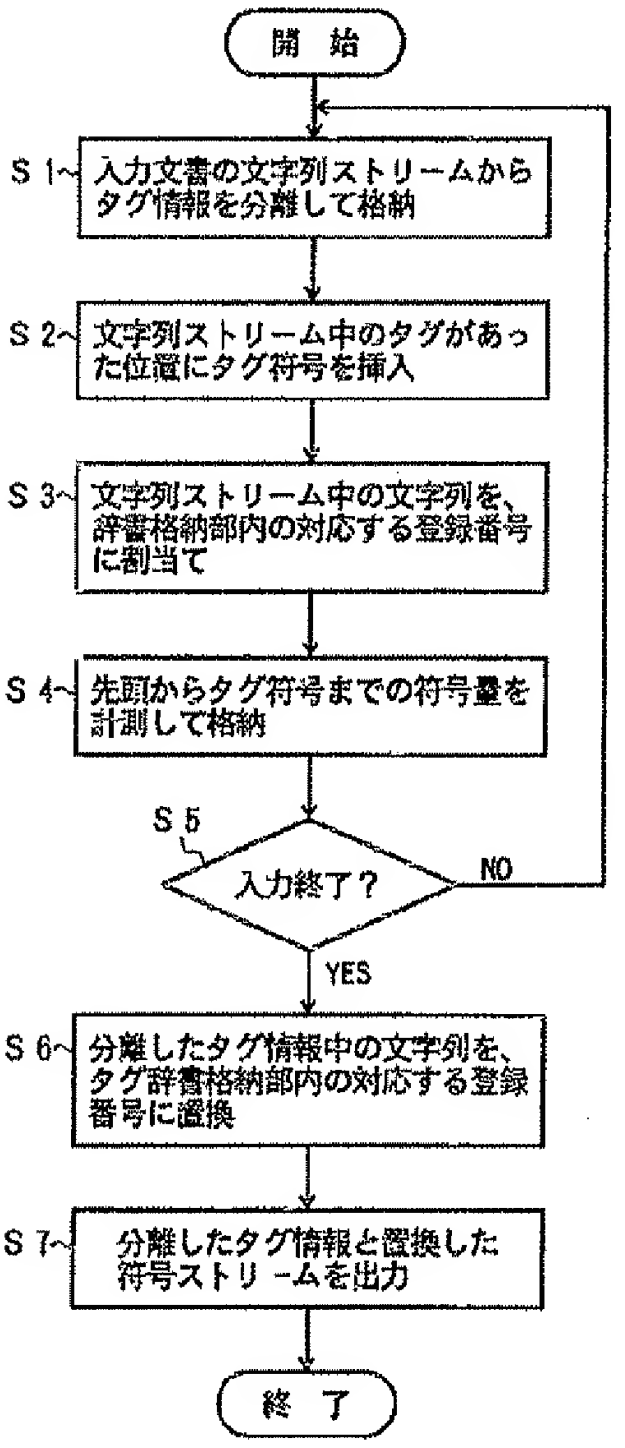
【図24】

図20のデータ圧縮装置からの符号ストリームを復元する本発明のデータ復元装置の第3実施形態のブロック図



【図23】

図20のデータ圧縮処理のフローチャート



【図27】

日本語文書を例にとったSGML文字ファイルの説明図

<TITLE>発明（考案）の明細書</TITLE>
<SECTION>発明（考案）の名称</SECTION>
　　データ圧縮・復元方法および装置並びに文書管理システム<P>
<SECTION>請求項目の範囲</SECTION>
<SUBSECTION>入力データを符号化するデータ圧縮装置において、</SUBSECTION>
<LIST>
　　<ITEM>符号表を保持する手段と
　　<ITEM>符号表を元に入力データを符号化する過程と
</LIST>を有することを特徴としたデータ圧縮方法。<P>
<SECTION>発明の詳細な説明</SECTION>
<SUBSECTION>産業上の利用分野</SUBSECTION>
<PARAGRAPH>近年、文字コード、ベクトル情報、画像など様々な種類のデータがコンピュータで扱われるようになっており、扱われるデータ量も急速に増加してきている。大量のデータを扱うときは、データの中の冗長な部分を省いてデータ量を圧縮することで、記憶容量を減らしたり、速く伝送したりできるようになる。
<PARAGRAPH>一方、最近では計算機で扱う文書の形式を統一する動きがある。これまで、計算機あるいはアプリケーションによってばらばらであった文書の形式をどんな計算機でも使用できるようにするものである。<TT>SGML（Standard Generalized Markup Language）</TT>は1986年にIS

【図29】

圧縮していないタグ情報の部分と圧縮した本文部分が混在したSGML文書圧縮データファイルの説明図

<TITLE>発明（考案）の明細書</TITLE><P>
<SECTION>発明（考案）の名称</SECTION>
275fe6afd8...　（圧縮領域）　...Cefc312903
<SECTION>請求項目の範囲</SECTION>
<SUBSECTION>入力データを符号化するデータ圧縮装置において、</SUBSECTION>
6ef208ca9d...　（圧縮領域）　...358eac0e86f8
<SECTION>発明の詳細な説明</SECTION>
<SUBSECTION>産業上の利用分野</SUBSECTION>
398ecad978...　（圧縮領域）　...ffd058e02b

【手続補正書】
【提出日】平成11年7月14日（1999．7．14）
【手続補正1】
【補正対象書類名】明細書
【補正対象項目名】特許請求の範囲
【補正方法】変更
【補正内容】
【特許請求の範囲】
【請求項1】タグを含む文書で構成された文字列ストリームから符号データを生成するデータ圧縮装置に於いて、
前記文字列ストリームから識別したタグを分離してタグ情報として出力するタグ情報分離部と、
前記タグ情報分離部でタグが分離された文字列ストリームの位置に識別のためにタグ符号を配置するタグ符号置換部と、
前記タグ符号置換部から出力されたタグ符号を含む文字列ストリームを符号化して符号ストリームを出力する文字列符号化部と、を有することを特徴とするデータ圧縮装置。
【請求項2】請求項1記載のデータ圧縮装置に於いて、前記タグ符号置換部は、タグが分離された文字列ストリームの位置に、所定の固定符号を前記タグ符号として配置することを特徴とするデータ圧縮装置。
【請求項3】請求項1記載のデータ圧縮装置に於いて、前記タグ符号置換部は、タグが分離された文字列ストリームの位置に、前記タグ情報分離部で分離されたタグの出現順序を示すタグ符号を配置することを特徴とするデータ圧縮装置。
【請求項4】請求項1記載のデータ圧縮装置に於いて、更に、
前記タグ情報分離部で分離されたタグ情報を格納するタグ情報格納部と、
前記文字列符号化部で生成された符号データを格納する符号格納部と、
前記タグ情報格納部に格納されたタグ情報と符号格納部

に格納された符号データを選択して出力する符号切替部と、を設けたことを特徴とするデータ圧縮装置。
【請求項５】請求項１記載のデータ圧縮装置に於いて、前記文字列符号化部は、
圧縮する際の処理単位となる文字列を登録した辞書を格納する辞書格納部と、
前記タグ符号置換部からの文字列ストリームの中の部分文字列と前記辞書格納部の登録文字列との比較により、前記登録文字列に一致する部分文字列を検出し、検出した部分文字列ごとに予め定められた符号を割り当てて出力する文字列比較部と、を備えたことを特徴とするデータ圧縮装置。
【請求項６】請求項１記載のデータ圧縮装置に於いて、更に、前記タグ情報分離部で分離したタグ情報を圧縮するタグ情報圧縮部を設けたことを特徴とするデータ圧縮装置。
【請求項７】請求項１記載のデータ圧縮装置に於いて、更に、
圧縮する際の処理単位となるタグ情報中のタグ文字列を登録した辞書を格納するタグ辞書格納部と、
前記タグ情報分離部で分離したタグ情報に含まれる文字列ストリームの部分文字列と前記タグ辞書格納部の登録文字列との比較により、前記登録文字列に一致する部分文字列を検出し、検出した部分文字列ごとに予め定められた符号を割り当てて出力するタグ文字列比較部と、を備えたことを特徴とするデータ圧縮装置。
【請求項８】請求項４記載のデータ圧縮装置に於いて、更に、前記文字列符号化部で生成した符号データの中のタグ位置を検出するタグ位置検出部を設け、前記タグ情報格納部に前記タグ情報分離部で分離したタグ情報と共に前記タグ位置検出部で検出したタグ位置の指定情報を格納したことを特徴とするデータ圧縮装置。
【請求項９】請求項８記載のデータ圧縮装置に於いて、前記タグ位置検出部は、文書先頭又は特定のタグからの符号量を検出して前記タグ情報格納部にタグ情報と共に格納したことを特徴とするデータ圧縮装置。
【請求項１０】タグを含む文書の文字列ストリームから分離したタグ情報と、分離したタグの位置にタグ符号を配置した文字列ストリームを符号化した符号データとを含む符号ストリームから文字列データを復元するデータ復元装置に於いて、
前記符号ストリームからタグ情報と符号データとを分離するタグ情報分離部と、　前記タグ情報分離部で分離したタグ情報を格納するタグ情報格納部と、
前記符号データから文字列及びタグ符号を含む文字列データを復元した後に、前記タグ符号をタグ情報格納部のタグ情報に置き換える文字列復元部と、を備えたことを特徴とするデータ復元装置。
【請求項１１】請求項１０記載のデータ復元装置に於いて、前記文字列復元部は、
復元する際の処理単位となる文字列の符号に対応した復元文字列を登録した辞書を格納する辞書格納部と、
前記符号ストリームから復元単位となる文字列の符号を分離して前記辞書格納部の参照で元の文字列を復元する文字列比較部と、
前記文字列比較部により復元したタグ符号を、前記タグ情報格納部のタグ情報に置き換える文字列置換部と、を備えたことを特徴とするデータ復元装置。
【請求項１２】請求項１０記載のデータ復元装置に於いて、更に、前記タグ情報格納部に格納されたタグ情報の圧縮データを復元するタグ情報復元部を設けたことを特徴とするデータ復元装置。
【請求項１３】請求項１０記載のデータ復元装置に於いて、更に、
復元する際の処理単位となるタグ文字列の符号に対応した復元文字列を登録した辞書を格納するタグ辞書格納部と、
前記タグ情報分離部により分離したタグ情報から復元単位となるタグ文字列の符号を分離し、前記辞書格納部の参照で元のタグ文字列を復元するタグ文字列比較部と、を備えたことを特徴とするデータ復元装置。
【請求項１４】タグを含む文書で構成された文字列ストリームから符号データを生成するデータ圧縮方法に於いて、
前記文字列ストリームから識別したタグを分離してタグ情報として出力するタグ情報分離過程と、
前記タグ情報分離過程でタグが分離された文字列ストリームの位置に識別のためにタグ符号を配置するタグ符号置換過程と、
前記タグ符号置換過程から出力されたタグ符号を含む文字列ストリームを符号化して符号ストリームを出力する文字列符号化過程と、を有することを特徴とするデータ圧縮方法。
【請求項１５】請求項１４記載のデータ圧縮方法に於いて、前記タグ符号置換過程は、タグが分離された文字列ストリームの位置に、所定の固定符号を前記タグ符号として配置することを特徴とするデータ圧縮方法。
【請求項１６】請求項１４記載のデータ圧縮方法に於いて、前記タグ符号置換過程は、タグが分離された文字列ストリームの位置に、前記タグ情報分離過程で分離されたタグの出現順序を示すタグ符号を配置することを特徴とするデータ圧縮方法。
【請求項１７】請求項１４記載のデータ圧縮方法に於いて、更に、
前記タグ情報分離過程で分離されたタグ情報を格納するタグ情報格納過程と、
前記文字列符号化過程で生成された符号データを格納する符号格納過程と、
前記タグ情報格納過程に格納されたタグ情報と符号格納過程に格納された符号データを選択して出力する符号切

替過程と、を設けたことを特徴とするデータ圧縮方法。
【請求項１８】請求項１４記載のデータ圧縮方法に於いて、前記文字列符号化過程は、
圧縮する際の処理単位となる文字列を登録した辞書を生成する辞書生成過程と、
前記タグ符号置換過程で得られた文字列ストリームの中の部分文字列と前記辞書の登録文字列との比較により、前記登録文字列に一致する部分文字列を検出し、検出した部分文字列ごとに予め定められた符号を割り当てて出力する文字列比較過程と、を備えたことを特徴とするデータ圧縮方法。
【請求項１９】請求項１４記載のデータ圧縮方法に於いて、更に、前記タグ情報分離過程で分離したタグ情報を圧縮するタグ情報圧縮過程を設けたことを特徴とするデータ圧縮方法。
【請求項２０】請求項１４記載のデータ圧縮方法に於いて、更に、
圧縮する際の処理単位となるタグ情報中のタグ文字列を登録した辞書を生成するタグ辞書生成過程と、
前記タグ情報分離過程で分離したタグ情報に含まれる文字列ストリームの部分文字列と前記タグ辞書の登録文字列との比較により、前記登録文字列に一致する部分文字列を検出し、検出した部分文字列ごとに予め定められた符号を割り当てて出力するタグ文字列比較過程と、を備えたことを特徴とするデータ圧縮方法。
【請求項２１】請求項１７記載のデータ圧縮方法に於いて、更に、前記文字列符号化過程で生成した符号データのタグ位置を検出するタグ位置検出過程を設け、前記タグ情報分離過程で分離したタグ情報と共に前記タグ位置検出過程で検出したタグ位置の指定情報を格納したことを特徴とするデータ圧縮方法。
【請求項２２】請求項２１記載のデータ圧縮方法に於いて、前記タグ位置検出過程は、文書先頭又は特定のタグからの符号量を検出して前記タグ情報格納過程で分離したタグ情報と共に格納することを特徴とするデータ圧縮方法。
【請求項２３】タグを含む文書の文字列ストリームから分離したタグ情報と、分離したタグの位置にタグ符号を配置した文字列ストリームを符号化した符号データとを含む符号ストリームから文字列データを復元するデータ復元方法に於いて、
前記タグ情報と符号データとを分離するタグ情報分離過程と、
前記タグ情報分離過程で分離したタグ情報を格納するタグ情報格納過程と、
前記符号データから文字列及びタグ符号を含む文字列ストリームを復元した後に、前記タグ符号を前記タグ情報格納過程で分離したタグ情報に置き換える文字列復元過程と、を備えたことを特徴とするデータ復元方法。
【請求項２４】請求項２３記載のデータ復元方法に於いて、前記文字列復元過程は、
復元する際の処理単位となる文字列の符号に対応した復元文字列を登録した辞書を生成する辞書生成過程と、
前記符号ストリームから復元単位となる文字列の符号を分離して前記辞書の参照で元の文字列を復元する文字列比較過程と、
前記文字列比較過程により復元したタグ符号を、前記ダク情報格納過程で分離したタグ情報に置き換える文字列置換過程と、を備えたことを特徴とするデータ復元方法。
【請求項２５】請求項２３記載のデータ復元方法に於いて、更に、前記タグ情報格納過程で格納されたタグ情報の圧縮データを復元するタグ情報復元過程を設けたことを特徴とするデータ復元方法。
【請求項２６】請求項２３記載のデータ復元方法に於いて、更に、
復元する際の処理単位となるタグ文字列の符号に対応した復元文字列を登録した辞書を生成するタグ辞書生成過程と、
前記タグ情報分離過程により分離したタグ情報から復元単位となるタグ文字列の符号を分離し、前記辞書の参照で元のタグ文字列を復元するタグ文字列比較過程と、を備えたことを特徴とするデータ復元方法。
【手続補正２】
【補正対象書類名】明細書
【補正対象項目名】０００５
【補正方法】変更
【補正内容】
【０００５】このような文書に構造の概念を取り入れた構造化文書の例としては、国際規格のＯＤＡ（ISO ８６１３：Open Document Architecture）や、ＳＧＭＬ（ISO8879：Standard Generalized Markup Language）の規格による構造化文書がある。またこのような構造文書を用いた文書処理方法は、例えば特開平５－１３５０５４号のものがある。
【手続補正３】
【補正対象書類名】明細書
【補正対象項目名】００１３
【補正方法】変更
【補正内容】
【００１３】ステップＳ１で同一の登録文字又は文字列が検索できなかった場合は、ステップＳ５で元の入力文字又は文字列をそのまま出力する。このような処理をステップＳ４で入力文字列がなくなるまで繰り返す。
【手続補正４】
【補正対象書類名】明細書
【補正対象項目名】００６０
【補正方法】変更
【補正内容】
【００６０】続いてステップＳ３で、タグ配置済みの文

字列ストリーム中の文字列に文字列符号化部１４に設けている文字列比較部１６で辞書格納部１８内の対応する登録番号を符号として割り当て、符号ストリーム２６を出力する。このステップＳ１～Ｓ３の処理を、ステップＳ４で文字列ストリームの入力が終了するまで繰り返す。

【手続補正５】
【補正対象書類名】明細書
【補正対象項目名】００６２
【補正方法】変更
【補正内容】
【００６２】図２の文字列符号化部１４に設けた文字列比較部１６は、辞書格納部１８の参照により、単語を構成する文字列ごとに予め定めた所定の文字列符号を割り当てる符号化を行う。

【手続補正６】
【補正対象書類名】明細書
【補正対象項目名】００６４
【補正方法】変更
【補正内容】
【００６４】文字列比較部１６は日本語文書データの文字列を先頭から順番に入力し、辞書格納部１８に予め登録されている単語単位の登録文字列と一致するか否かを検出する。文字列比較部１６で入力文字列に一致する登録文字列が検出されると、辞書格納部１８の一致検出された登録文字列に対応して予め登録されている文字列符号を読み出して割り当て、この文字列符号を出力する。

【手続補正７】
【補正対象書類名】明細書
【補正対象項目名】００６５
【補正方法】変更
【補正内容】
【００６５】ここで日本語文書データの文字列を単語単位に文字列符号に変換するための辞書格納部１８を説明する。

【手続補正８】
【補正対象書類名】明細書
【補正対象項目名】００６９
【補正方法】変更
【補正内容】
【００６９】図１０は、図２の辞書格納部１８の辞書構造の実施形態である。図２の辞書格納部１５に格納された辞書は、先頭文字格納部４０と従属文字列格納部４２の２階層構造を備える。先頭文字格納部４０は、日本語文字「あ，い，う，え，お・・・」の文字コードをインデックスとしており、日本語の文字コードは２バイトデータであることから、文字コード４４としては、１６進数で「０ｘ０００」から「０ｘＦＦＦＦ」の１３１，０７２種類の格納位置が割り当てられる。

$$K=(N\cdot X-A1)/M$$

但し、Ｘ　：従属文字列格納部４２の位置アドレス
Ｎ　：一致検出された従属文字列の番号（１，２，３，・・Ｎ）
Ａ１：従属文字列格納部の開始アドレス
Ｍ　：従属文字列格納部の格納バイト長

【手続補正９】
【補正対象書類名】明細書
【補正対象項目名】００７０
【補正方法】変更
【補正内容】
【００７０】この文字コード４４は、図２の文字列比較部１６で読み込んだ先頭文字を使用して、対応する文字コードの位置にアクセスする。文字コード４４に続いては先頭アドレス４６が格納される。先頭アドレス４６は、例えば文字コード４４の先頭文字「あ」を例にとると、先頭文字「あ」に続く従属文字列を格納した従属文字列格納部４２の先頭アドレス「Ａ１」を指定している。続いて従属文字列の個数４８が設けられる。例えば先頭文字「あ」にあっては、従属文字列個数４８としてＮ１＝４個が格納されている。

【手続補正１０】
【補正対象書類名】明細書
【補正対象項目名】００７１
【補正方法】変更
【補正内容】
【００７１】従属文字列格納部４２は、先頭文字格納部４０の先頭文字の文字コード４４に対応して格納された先頭アドレス４６で先頭位置が指定され、この先頭位置から従属文字列格納部４２で指定された個数の格納位置に従属文字列が格納されている。例えば先頭文字「あ」に対応した先頭アドレス４６のアドレスＡ１から従属文字列個数４８のＮ１＝４個となる４つの格納位置が、対象とする従属文字列格納領域として指定される。

【手続補正１１】
【補正対象書類名】明細書
【補正対象項目名】００７４
【補正方法】変更
【補正内容】
【００７４】ここで図１０の従属文字列格納部４２に文字列コード３４は、単語個数に基づき１番から１３６，４８６番まで予め１７ビットの文字列コードが割り当てられており、図１０のように格納した場合の文字列コード（文字列符号）Ｋと位置アドレスＸとの関係は、次式で表すことができる。

（１）

ここで、従属文字列格納部４２の格納バイト長Ｍは、従属文字列の長さ５０、従属文字列５２、及び文字列コード５４の合計長であることから、例えば次式で表すことができる。

格納バイト長M＝　長さ＋文字コード列＋文字列コード　（２）
＝3bit＋　96bit　＋　17bit
＝116bit
＝15byte

ここでは格納可能な従属文字列を最大６文字とすることで、従属文字列５２に９６ビットを割り当てた場合を例にとっている。もちろん、従属文字列の文字数は平均で２．８文字であることから、３文字（４８ビット）以上とすれば十分な圧縮効果が得られる。この場合には、従属文字列格納部の１つの格納領域の格納バイト長MはM＝１２バイトになる。

【手続補正１２】
【補正対象書類名】明細書
【補正対象項目名】００７６
【補正方法】変更
【補正内容】
【００７６】図１１は図１０の辞書構造を持った辞書格納部１８による図２の文字列比較部１６による符号化処理のフローチャートである。

【手続補正１３】
【補正対象書類名】明細書
【補正対象項目名】００７９
【補正方法】変更
【補正内容】
【００７９】登録した従属文字列と一致すれば、ステップＳ８に進み、次の文字列コード５４を読み出し、一致検出した文字列に文字列比較部１６で割り当てて出力し、次のステップＳ９で先頭文字位置Ｐを従属文字列の文字数Ｌだけ移動した位置Ｐにポインタを更新する。そしてステップＳ１２で非圧縮データの処理が済んでいなければ、再びステップＳ２に戻り、更新した先頭文字位置Ｐについて同様な処理を繰り返す。

【手続補正１４】
【補正対象書類名】明細書
【補正対象項目名】００８２
【補正方法】変更
【補正内容】
【００８２】図１２は、図２のデータ圧縮装置から出力された符号ストリーム２６及びタグ情報ストリーム２８で構成される符号ストリームから文字列ストリームを復元するためのデータ復元装置の第１実施形態のブロック図である。

【手続補正１５】
【補正対象書類名】明細書
【補正対象項目名】００８３
【補正方法】変更
【補正内容】
【００８３】このデータ復元装置は、タグ情報分離部６０、タグ情報格納部６２、文字列復元部６４で構成される。文字列復元部６４は、符号列比較部６６、辞書格納部６５及び文字列置換部６８を備える。

【手続補正１６】
【補正対象書類名】明細書
【補正対象項目名】００８４
【補正方法】変更
【補正内容】
【００８４】タグ情報分離部６０は図２のデータ圧縮装置側から送られてきた符号ストリーム５６を入力し、タグ情報と符号データとに分離し、タグ情報はタグ情報格納部６２に格納し、符号データは符号ストリーム５８として文字列復元部６４に出力する。

【手続補正１７】
【補正対象書類名】明細書
【補正対象項目名】００８５
【補正方法】変更
【補正内容】
【００８５】文字列復元部６４は符号列比較部６６で辞書格納部６５を用いて符号データから文字列及びタグ符号を復元した後に、文字列置換部６８においてタグ符号をタグ情報格納部６２に格納しているタグ情報に置き換えて、復元した文字列ストリーム７０を出力する。

【手続補正１８】
【補正対象書類名】明細書
【補正対象項目名】００８６
【補正方法】変更
【補正内容】
【００８６】図１３は、図１２のデータ復元装置の復元処理のフローチャートである。まずステップＳ１で、タグ情報分離部６０が入力文書に対応した符号ストリーム５６からタグ情報を分離してタグ情報格納部６２に格納する。次にステップＳ２で、タグ情報が分離された符号ストリーム５６の中の符号列を辞書格納部６５内の登録番号と比較照合し、一致する登録番号で格納している文字または文字列に変換する。

【手続補正１９】
【補正対象書類名】明細書
【補正対象項目名】００８８
【補正方法】変更
【補正内容】
【００８８】図１２の文字列復元部６４に設けた符号列比較部６６は、辞書格納部６５の参照により、図３のデータ圧縮装置で符号化された符号列ストリームから元の文字列を復元する。

【手続補正２０】

【補正対象書類名】明細書
【補正対象項目名】００８９
【補正方法】変更
【補正内容】
【００８９】図１４は、図１２の文字列辞書格納部６５の辞書構造である。この文字列辞書格納部６５にあって

$$\text{格納バイト長M} = \text{先頭文字} + \text{長さ} + \text{文字コード列} \quad (6)$$
$$= 16\text{bit} + 3\text{bit} + 96\text{bit}$$
$$= 115\text{ビット}$$
$$= 15\text{byte}$$

から判明しているため、次式から文字列コードKに対応

$$X = M \cdot K + A1$$

但し、K　：文字列コード
Ａ１：文字列格納位置の開始アドレス
Ｍ　：格納バイト長
このようにして分離した文字列コードＫから辞書格納位置を示す位置アドレスＸを求めて参照することで、対応する先頭文字及び従属文字列を組み合わせた文字列を復元することができる。
【手続補正２１】
【補正対象書類名】明細書
【補正対象項目名】００９３
【補正方法】変更
【補正内容】
【００９３】タグ情報格納部７８にはタグ情報分離部１０により文字列ストリーム２０から分離されたタグ情報が格納される。これによってタグ情報格納部７８には、例えば図６のようなタグ情報ファイル３６が格納される。また符号格納部８０は文字列符号化部１４に設けられており、タグ符号置換部１２により分離したタグ情報にタグ情報を挿入したタグ置換済み文字列ストリーム２２につき、図１１の符号化処理により生成された符号データが格納される。
【手続補正２２】
【補正対象書類名】明細書
【補正対象項目名】００９７
【補正方法】変更
【補正内容】
【００９７】文字列ストリームの入力が終了するとステップＳ５に進み、分離したタグ情報とタグ符号に変換して符号化した符号ストリームを、例えばタグ情報格納部７８と符号格納部８０から順番に読み出して符号列ストリーム８４として出力する。図１５のデータ圧縮装置から出力された符号列ストリーム８４は、図１２に示したデータ復元装置に入力することで文字列ストリームを復元することができる。
【手続補正２３】
【補正対象書類名】明細書
【補正対象項目名】０１０１
【補正内容】
【０１０１】タグ情報分離部１０、タグ符号置換部１２、文字列符号化部１４は、図１５の第２実施形態と同じである。
【手続補正２４】
【補正対象書類名】明細書
【補正対象項目名】０１０４
【補正方法】変更
【補正内容】
【０１０４】図１９は、図１７のデータ圧縮装置から出力された符号ストリーム９０から文字列ストリームを復元する本発明のデータ復元装置の第２実施形態である。このデータ復元装置は、図１２の第１実施形態に更に圧縮タグ格納部９２とタグ情報復元部９４を設けている。
【手続補正２５】
【補正対象書類名】明細書
【補正対象項目名】０１０５
【補正方法】変更
【補正内容】
【０１０５】タグ情報分離部９２は、入力する符号ストリーム９０に含まれる圧縮タグ情報を分離して圧縮タグ格納部６２に格納する。圧縮タグ格納部９２に格納された圧縮タグ情報はタグ情報復元部９２により復元され、タグ情報格納部６２に格納される。タグ情報復元部９２はデータ圧縮側のＬＺ７７，ＬＺ７８，算術復号化に対応した復元アルゴリズムを実行する。それ以外の構成は図１５と同じになる。
【手続補正２６】
【補正対象書類名】明細書
【補正対象項目名】０１０７
【補正方法】変更
【補正内容】
【０１０７】図２０において、タグ情報分離部１０、タグ符号置換部１２、文字列比較部１６、辞書格納部１８を備えた文字列符号化部１４、タグ情報格納部７８及び符号切替部８２は、図１５の第２実施形態と同じであ

は、先頭文字７２、従属文字列長さ７４及び従属文字列７６を、図１０の辞書構造に示した従属文字列格納部４２の１７ビットの文字列コード５４の順番に格納している。このため符号列比較部６６にあっては、復元に使用する従属文字列格納部４２の格納バイト長Mが

から判明しているため、次式から文字列コードKに対応した位置アドレスXを算出することができる。

$$X = M \cdot K + A1 \quad (7)$$

る。これに加えて図２０の第４実施形態にあっては、新たにタグ文字列比較部９７、タグ辞書格納部９６及び符号量計測部９８を設けている。

【手続補正２７】
【補正対象書類名】明細書
【補正対象項目名】０１０８
【補正方法】変更
【補正内容】
【０１０８】タグ文字列比較部９７とタグ辞書格納部９６は、タグ情報分離部１０で分離したタグ情報に含まれる日本語文字列ストリームを文字列符号化部１４と同様な符号化アルゴリズムで符号化してタグ情報を圧縮する。このため、タグ情報格納部９８の辞書構成は図１０と同じであり、先頭文字及び従属文字としてタグ情報に使用する日本語文字列が使用されている。またタグ文字列の符号化処理は図１１のフローチャートに従って行う。

【手続補正２８】
【補正対象書類名】明細書
【補正対象項目名】０１１１
【補正方法】変更
【補正内容】
【０１１１】図２１は、図２０の第４実施形態における圧縮処理の説明図である。ＳＧＭＬ日本語文書ファイル３５の内容となる文字列ストリームを入力して、タグ情報の分離によるタグ情報ファイル３６の生成及びタグ情報をタグ符号に置換したタグ置換済み日本語文書ファイル３８の生成は、図１５の第２実施形態と同じである。

【手続補正２９】
【補正対象書類名】明細書
【補正対象項目名】０１１３
【補正方法】変更
【補正内容】
【０１１３】図２２は、タグ情報格納部７８に格納されたタグ情報ファイルの具体例であり、図２７に示したＳＧＭＬ日本語文書ファイルから分離したタグ情報を例にとっている。このタグ情報ファイル３６には、左側のインデックス０１～１３に対応した各タグに対応して、右側に図２１のタグ置換済み日本語ファイル３８の文字列データの符号データの先頭からの符号量（バイト量）ＤＬ１～ＤＬ１３が位置指定情報１０６としてそれぞれ格納されている。

【手続補正３０】
【補正対象書類名】明細書
【補正対象項目名】０１１６
【補正方法】変更
【補正内容】
【０１１６】このようなステップＳ１～Ｓ４の処理を、ステップＳ５で文字列ストリームの入力が終了するまで繰り返す。文字列ストリーム２０の入力が終了すると、ステップＳ６で、タグ情報格納部７８に分離して格納しているタグ情報中の文字列をタグ辞書格納部９６内の辞書の対応するブロック番号に変換して符号データとする符号化処理をタグ文字列比較部９７で行い、タグ情報格納部７８に格納する。その結果、タグ情報格納部７８の格納内容は図２２の圧縮タグ情報ファイル３６のようになる。

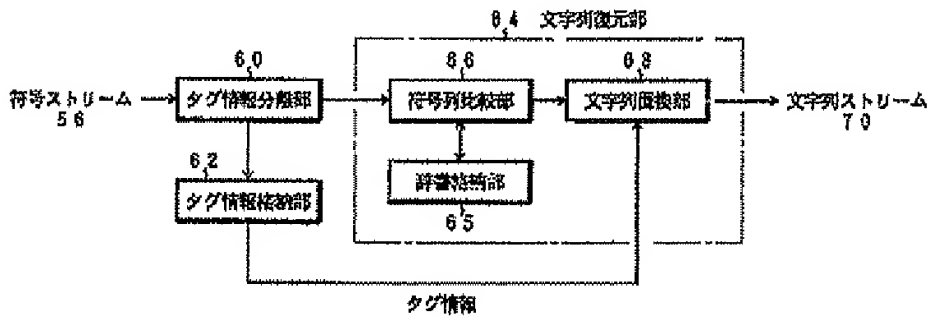
【手続補正３１】
【補正対象書類名】明細書
【補正対象項目名】０１１７
【補正方法】変更
【補正内容】
【０１１７】最後にステップＳ７で、タグ情報格納部７８に分離して符号化した符号量付きのタグ情報と符号格納部８０に格納した符号データを符号切替部８２より例えば順番に選択出力し、符号ストリーム１００として外部に供給する。

【手続補正３２】
【補正対象書類名】明細書
【補正対象項目名】０１２３
【補正方法】変更
【補正内容】
【０１２３】一方、タグ情報分離部６０は、圧縮タグ情報ストリームに続いて送られてくる文書本文の符号ストリームを文字列復元部６４に供給し、符号列比較部６６で取り出した符号による辞書格納部６５の辞書番号の参照で対応する文字または文字列を復元し、文字列置換部６８に出力する。

【手続補正３３】
【補正対象書類名】明細書
【補正対象項目名】０１２４
【補正方法】変更
【補正内容】
【０１２４】文字列置換部６８は、復元した文字列の中のタグ符号を認識し、その出現順に従ってタグ情報格納部６２に格納している復元済みのタグ情報を格納順に取り出し、タグ符号と置換し、復元した文字列ストリームを出力する。

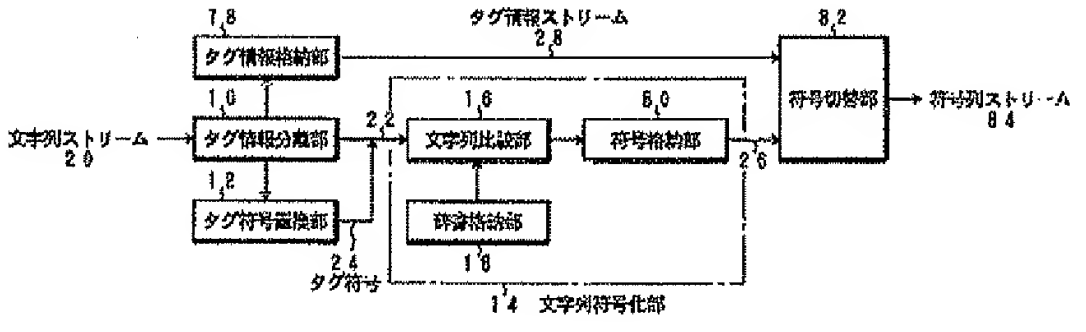
【手続補正３４】
【補正対象書類名】図面
【補正対象項目名】図１２
【補正方法】変更
【補正内容】
【図１２】

図2のデータ圧縮装置からの符号ストリームを復元する本発明のデータ復元装置の第1実施形態のブロック図



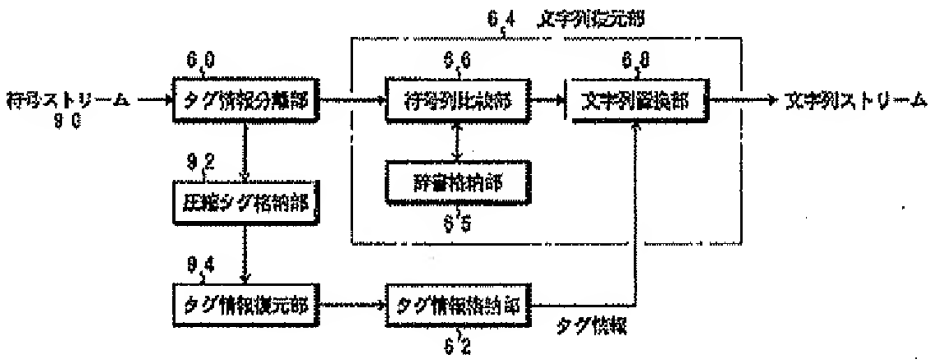
【手続補正３５】
【補正対象書類名】図面
【補正対象項目名】図１５
【補正方法】変更
【補正内容】
【図１５】

本発明のデータ圧縮装置の第２実施形態のブロック図



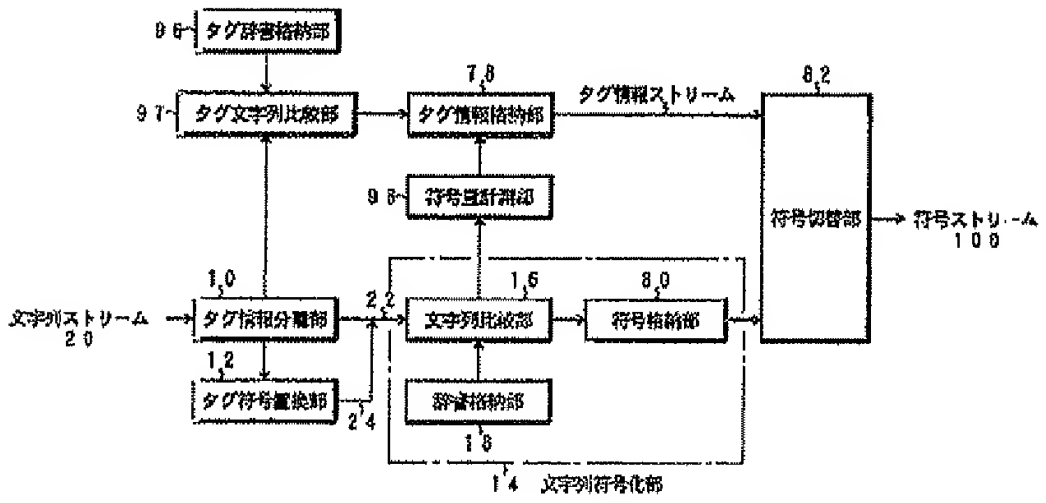
【手続補正３６】
【補正対象書類名】図面
【補正対象項目名】図１９
【補正方法】変更
【補正内容】
【図１９】

図17のデータ圧縮装置からの符号ストリームを復元する本発明のデータ復元装置の第２実施形態のブロック図



【手続補正３７】
【補正対象書類名】図面
【補正対象項目名】図２０
【補正方法】変更
【補正内容】
【図２０】

本発明のデータ圧縮装置の第４実施形態のブロック図



フロントページの続き

(72)発明者　佐藤　宣子
神奈川県川崎市中原区上小田中４丁目１番
１号　富士通株式会社内